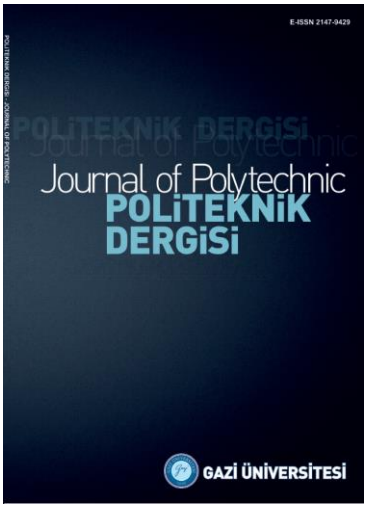

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

CLARIVATE ANALYTICS
EMERGING SOURCES CITATION INDEX
INDEXED

A hybrid spam detection framework for social networks

Sosyal ağlar için bir hibrit spam algılama modeli

***Authors (Yazarlar)**: Oğuzhan ÇITLAK[1], Murat DÖRTERLER[2], İbrahim Alper DOĞRU[3]*

ORCID[1]: 0000-0001-9545-2795

ORCID[2]: 0000-0003-1127-515X

ORCID[3]: 0000-0001-9324-7157

ERKEN GÖRÜNÜM

To cite to this article (Bu makaleye şu şekilde atıfta bulunabilirsiniz): Çıtlak O., Dörterler M. ve Doğru İ. A., "A hybrid spam detection framework for social networks", ***Journal of Polytechnic***, *(*): *, (*).

To link to this article (Erişim linki): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.933785**

A Hybrid Spam Detection Framework for Social Networks

Highlights

- *Obtaining dynamic dataset over social networks.*
- *Using the machine learning model in the dataset.*
- *Using the Short Link analysis model on the dataset.*
- *Using the text analysis model on the dataset.*
- *Detection of spam on social networks.*

Graphical Abstract

The data set is obtained through social networks. This dataset is evaluated with an application developed simultaneously with three different spam detection models.

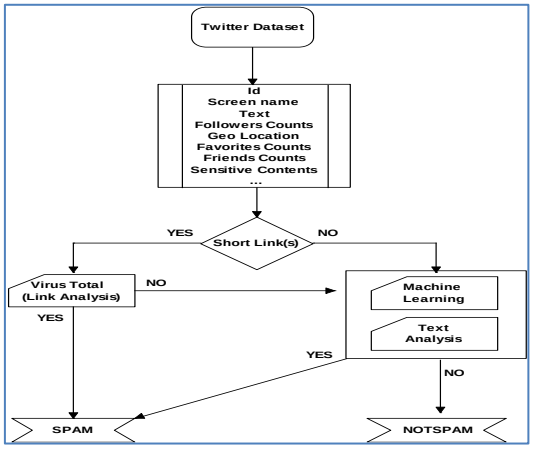

Fig. Application working diagram

Aim

In this study, it is tryied to bring together three spam detection models and use them at the same time. It is aimed to detect spam accounts on social networks and to contribute to the spam detection policies applied by social networks.

Design & Methodology

An application has been developed in which the proposed hybrid spam detection model is used together. The dataset used is run on this application and its results are evaluated.

Originality

It is applied of machine learning, link analysis and text analysis spam detection model that could be evaluated differently from each other and simultaneously running on the dataset obtained over the social network.

Findings

The success achieved is compared with other studies in the literature. 95.69 % success rate is calculated. A more successful result is obtained compared to previous similar studies.

Conclusion

In many studies in the literature, spam detection on social networks can be done on datasets. However, in future studies, spam detection in social networks will be possible immediately without a dataset. In addition to this, different spam detection models could achieve higher success when used together.

Declaration of Ethical Standards

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

A Hybrid Spam Detection Framework for Social Networks

Research Article / Araştırma Makalesi

Oğuzhan ÇITLAK[1*], Murat DÖRTERLER[2], İbrahim Alper DOĞRU[2]

[1]Department of Computer Engineering, Graduate School of Natural and Applied Sciences, Gazi University, Ankara, Turkey
[2]Department of Computer Engineering, Faculty of Technology, Gazi University, Ankara, Turkey

(Geliş/Received : 06.05.2021 ; Kabul/Accepted : 04.02.2022 ; Erken Görünüm/Early View : 10.03.2022)

ABSTRACT

The widespread use of social networks has caused these platforms to become the target of malicious people. Although social networks have their own spam detection systems, these systems sometimes may not prevent spams in their social networks. Spam contents and messages threaten the security and performance of users of these networks. A spam account detection framework based on three components is proposed in this study. Short link analysis, machine learning and text analysis are the components used together in the proposed framework. First, a dataset was created for this purpose and the attributes of spam accounts were determined. Later, the hyperlinks in the messages in this dataset were analyzed through link analysis component. The machine learning component was modelled through attributes. Moreover, the messages of the social network users were analyzed through text analysis method. A web-based application of the proposed model was put into practice. As a result of the experimental studies carried out thanks to the framework, it was determined that the proposed framework showed a performance of 95.69 %. The success of this article was calculated according to the F-measure and precision evaluation metrics under the influence of sensitive content rate. It is aimed to detect spam accounts on social network and the spam detection policy of these networks is intended to support.

Keywords: Social networks, spam detection, short link analysis, machine learning, text analysis.

Sosyal Ağlar için Bir Hibrit Spam Algılama Modeli

ÖZ

Sosyal ağların yaygınlaşması bu platformların kötü niyetli kişilerin hedefi haline gelmesine neden olmaktadır. Sosyal ağların kendi spam tespit sistemleri olmasına rağmen, bu sistemler bazen sosyal ağlarındaki spamları engelleyememektedir. Spam içerikler ve mesajlar, sosyal ağ kullanıcılarının güvenliğini ve performansını tehdit etmektedir. Bu çalışmada, üç bileşene dayalı bir spam hesap tespit modeli önerilmektedir. Kısa bağlantı analizi, makine öğrenmesi ve metin analizi önerilen modelde birlikte kullanılan bileşenlerdir. Bu amaçla, öncelikle bir veri seti oluşturulmuştur ve spam hesapların özellikleri belirlenmiştir. Sonra, bu veri setindeki mesajlarda yer alan hyperlinkler link analizi bileşeni ile analiz edilmektedir. Makine öğrenimi bileşeni, önceden belirlenen özniteliklere göre modellenmektedir. Ayrıca, sosyal ağ kullanıcılarının mesajları metin analizi yöntemi ilede analiz edilmektedir. Önerilen modelin web tabanlı bir uygulaması hayata geçirilmektedir. Önerilen model sayesinde yapılan deneysel çalışmalar sonucunda, önerilen modelin %95.69 oranında doğru performans gösterdiği tespit edilmektedir. Bu makalenin başarısının hesağlanmasında, hassas içerik oranının etkisi ile F puanı ve kesinlik değerlendirme metriklerine göre hesaplanmaktadır. Bu çalışmada, sosyal ağlardaki spam hesapların tespit edilmesi ve bu ağların spam tespit politikasının desteklenmesi amaçlanmaktadır.

Anahtar Kelimeler: Sosyal ağlar, spam tespiti, kısa bağlantı analizi, makine öğrenimi, metin analizi.

ERKEN GÖRÜNÜM

1. INTRODUCTION

Social platforms in which individuals can easily express themselves and share information have had a very important place in our lives [1]. Social networks such as Twitter take a significant part in social communication and communication among people [2]. Spam contents and messages have bad effects on the functionality of social networks, and also, they threaten the security and performance of users [3]. In this study, a new learning model based on three components is proposed for the detection of spam accounts in social networks. These components are link analysis, machine learning and text analysis. In the link analysis, the links in the messages sent by Twitter, which is one of the most used platforms in social services [3, 4] are examined. Virus Total tool was used in the link analysis part [5, 6]. Whether these links are in a repository in which malicious websites are constantly updated is checked. If the analyzed account shares a malicious link, this account can be called a spam account according to the spam policy of Twitter [7]. Namely, if a user is sharing malicious links, this account can be considered as a spam account. With this respect, a dataset created from Twitter is used for this study. The attributes of the machine learning components were created by means of examining the features of spam and non-spam accounts in the dataset. The crowdsourcing method was used for labelling in the analysis of Twitter

*Sorumlu Yazar (Corresponding Author)
e-posta : oguzhan.citlak@gazi.edu.tr

accounts used in machine learning attributes [8, 9, 10]. It is a crowdsourcing method that can be briefly defined as an online distributed problem solving and production model [9]. In this method, support is requested from people on the internet for the realization of a particular project [10]. The attributes of the analyzed account were evaluated by considering the machine learning method [9]. Whether the accounts were spam or not was endeavored to be determined based on their attributes. Finally, in the text analysis method, however, the texts in the messages of Twitter users were examined [10, 11]. The sensitive content words used by spam accounts in their messages are predetermined. Within the spam account messages, whether or not the words with sensitive content are communicated is checked. Therefore, whether an account sending messages on Twitter was spam or not was decided by evaluating the resultants of these components. In this article, a framework had been created from spam analysis models that are frequently mentioned in the literature [3, 12, -, 21]. Spam analysis models generally tried to get better results in the same dataset (some of them was out of date). In this study, a spam analysis framework model is proposed in the dynamic dataset obtained via Twitter API [22, 23]. The remainder of this study is organized as follows: the second section provides information about previous studies and the most commonly used spam detection methods. The third section introduces the proposed hybrid framework. The fourth section shows the results of the experimental studies and the comparison of the performance of the model with the other studies. In the last section, the findings are evaluated, and the success of the model is calculated.

2. RELATED STUDIES in LITERATURE

Many studies have been conducted to identify and remove spam and spam accounts [3, 15, 16, 18, 19, 24, 25]. The best precaution to be taken against these threats, which use social media extensively, is to know the ways in which spammers threaten users and to take personal precautions against them [26, 27]. Facebook, Twitter, LinkedIn, WhatsApp and Instagram are the most common social networks on the Internet that are used for different purposes [2, 4, 28, 29]. When classifying spam detection methods on social networks, the methods that are most up- to-date and most commonly used are taken into consideration [3, 24, 25, 27, 30].

Anomaly spam detection method is based on the behavior the users exhibit [12]. In the anomaly detection method detects unexpected situations in a data. In fact, unexpected behavior is when a data does not perform its expected behavior [12]. It is important to know normal behavior and to distinguish anomaly behaviors from the normal ones. The observation of behaviors is based on normal behavior patterns. Suspicious behaviors are compared with normal behaviors, and this behavior is detected and differentiated [12].

URL tracking system blacklists redirect URLs to block them using web-based Domain Genetic Algorithm - DGA[13]. In the studies of Akiyama et al. used short link method. More than 100 000 malicious redirect URLs were collected from 776 different websites [13]. The majority of click- fraud attackers use URL redirection. The most practical measure against malicious URL redirects is that security. Moreover, the infrastructure of web-based attacks is broken down, and these security or network operators are prevented [31]. In addition to this, spam account on social networks may also try to obtain the usernames and passwords of real users on network via sending malicious web link [6, 7, 31, -, 33].

In a study conducted by Fernandes et al. in 2015, compare and contrast methods were used and a similar F1 accuracy score of 90% was achieved [14]. However, this F1 accuracy score had some problems in classifying abnormal behaviors of real users. To avoid this, a secondary classification method was employed and F1 accuracy with an average of 74% was achieved. This accuracy is achieved by reducing the size of the property area and by using category balancing that is formed by gradual feature selection and individual control of category results [14]. A common example of spam used in social networks is deceptive spam. These spammers often propagate deceptive and misleading information and content [14]. The users are redirected to malicious sites or addresses with fake messages that appeal to the user, which are remarkable and apparently contain no harmful elements. A regional analysis of the responses given to these deceptive messages and which sites they are directed to as well as what kind of information is requested from the users is made and spam detection is thus provided [14].Social accounts are the profiles that internet users have on social websites. Your profile on Twitter can be considered as your social account [25].

The relationship among the used social accounts is analyzed in follow and follower comparison method. In a study by Wang in which this method was used a new method was developed through an API [22] provided by Twitter and a web browser [15]. A total of 25 000 users, 500 000 Tweets and 49 million followers and friends were collected from the publicly available data on Twitter. Naive Bayesian classification algorithm [34] was applied within the machine learning system for distinguishing suspicious behaviors from the normal ones. The dataset was analyzed, and the performance of the detection system was compared to the traditional evaluation matrixes and various classification methods [15]. The obtained results demonstrate that F-Scale matrix of Naive Bayesian classification algorithm has the best overall performance. When the entire trained dataset is tested; the result shows that the spam detection system can achieve an accuracy value of 89% [15].

In a study carried out by Romo and Araujo employing the Trend topics analysis method; spam in real time was detected by using language as a primary tool [16]. For the experimental study, they collected a large dataset with 34 000 trend-topics and 20 million tweets. In addition to this set, they created a table of specific features that were not altered but reduced by spammers. They also developed a

ERKENCODUM

machine learning system that had some features to be combined with other features to analyses the characteristics of spams in social networks [10, 24, 29, 35]. Moreover, they conducted a comprehensive evaluation of the established performance of the system based on the F-Scale matrix. It was shown the same level as the most advanced technology systems based on the detection of spam accounts. It is seen, as a result of this evaluation, that the system proposed is useful for detection of spam in trend-topics by means of analyzing tweets instead of user accounts [16].

In a study carried out by Liu et al., they demonstrated that the unstable distribution between spam and non- spam classes had a significant effect on spam detection rate [17]. To solve this problem, Fuzzy-based Information Decomposition Oversampling (FOS) algorithm was implemented [36]. They proposed a new fuzzy- based method that produced a synthetic dataset from limited observed samples. They also developed a community learning approach, and thanks to this approach they learned thanks to a more accurate classifier from data that appeared to be unstable in three stages. In this method, the class distribution in the unstable dataset was initially arranged. Secondly, a classification model was built on each of a reclassified dataset. At the final stage, a majority voting system was developed to combine the results from all classification models. For evaluation aims, the obtained results from the experiments carried out on Twitter data demonstrated that the proposed learning approach could significantly increase the spam detection rate in the unstable datasets [37]. Conventional detection methods based on the attributes of accounts or messages spend a considerable amount of time while collecting such information before running detection algorithms [38].

In their study, Lee and Kim proposed a new detection scheme to filter around the time of creating potentially malicious account groups [18]. In similar algorithms, the differences between the created account names “algorithmically” and actual account names are used to identify malicious accounts [26]. For the accounts created in a short period of time, they applied a separate classification algorithm to classify group accounts and malicious account clusters that shared similar username properties. They used 4.7 million user accounts collected from Twitter as the dataset. Even though this scheme is based on user account names and duration of creation, it achieves an acceptable accuracy value. This method has a structure that can be used as a quick filter to perform a detailed analysis against malicious account groups. In social networks, one of the most common situations is the presence of a large number of incoming bulk messages. Although these undesired bulk messages can be effectively distinguished by existing spam filters, message instances are modified to mislead spam filters [18].

In the study made by Miller et al., they defined spam detection as an anomaly detection problem rather than a spam classification problem [19]. They used the attributes of the user information and tweet texts from previous studies as datasets. To make designation of the spam identity easier, they employed the two stream clustering algorithms, namely “StreamKM++” and “DenStream by using the flow characteristics of tweets effectively. Both algorithms are used to group Tweeter users and mark anomalous names as spammers. When these algorithms are tested separately, it is seen that they perform well. 99% recall and 6.4% false positive rates were obtained through StreamKM++ and 99% recall and 2.8% false positive rates were obtained through DenStream [26]. When these algorithms are used together, it is seen that the system detects only 2.2% of normal users incorrectly while detecting spam users, and it is capable of detecting all spam senders correctly [19].

Maio et al. [20] created time-sensitive adaptive tweet sequences with a deep learning model in their study. Sequences; time, place, semantics, quality etc. it is affected by various situations. In the proposed model, tweet sequences depend on how to be the twitter account logged in, the user's interests and the time they log in. The sensitivity of the proposed model was evaluated with metrics Mean Average Precision (MAP) [39] and Normalized Discount Cumulative Gain (NDCG) [40]. Tweets are modeled as numerical feature vectors using the Word2Vec tool [41]. Comparative Neural Network (CmpNN) deep learning model has been chosen as the proposed method. Twitter dataset was used as the data set. The dataset was obtained between (26/01/2017 - 12/02/2017). The first 5 000 user accounts were removed from the data set and their behavior was analyzed. Approximately 93.8% accuracy rate had been achieved in the proposed model [20].

Chatterjee et al. [21] conducted a study on Deep Learning and Understanding Text Emotions Using Big Data. Emotions are physiological states produced in humans in response to internal and external events. As humans, "Why don't you text me at all!" in this message it can be interpreted as either a sad or angry emotion, and there is the same uncertainty for machines. The lack of facial expressions and voice modulation makes detecting emotions in text a difficult problem. In this study, a new Deep Learning based approach is proposed in textual dialogs Happy, Sad and Furious to detect emotions. The study uses Long Short-Term Memory (LSTM) [42], which uses a word comparison matrix in two different layers. The first layer uses a semantic word embedding. The other layer uses a sentiment word embedding. Word2Vec [41], Glove [43] and FastText [44] were used in semantic word analysis. 17.62 million tweet conversation pairs were used in the study. The Microsoft Cognitive Toolkit [45] has been used for training and maximum prediction accuracy. Approximately 93.2% accuracy rate had been achieved in the proposed model. Combining both semantic and emotion-based presentations was at the core of the approach to more accurate emotion detection [21]. The next part of this study showed the spam detection framework model. Link analysis, Machine Learning and Text Analysis methods

ERKENCOPY.COM

are the three main parts of proposed framework for social networks.

3. MATERIAL and METHOD

3.1. Hybrid Spam Detection Framework for Social Networks

In this study, a learning-based spam account detection framework has been proposed. This proposed model can be used in Twitter, which has an important place in social networks. This study may contribute to Twitter's spam policy to detect spam account. In the literature, there are many studies using datasets for spam detection related to Twitter [3, 9, 11, 15, 16, 19].

3.1.1. About Dataset

A dataset was created to develop and measured for our model. This section describes the components of the proposed framework. To measure the performance of the model, a comprehensive dataset was created over Twitter. Date Capture Start Date and Date Capture End Date indicate the date range in which the dataset mentioned in Table 3.1 was captured. Number of Users represents the total number of users in the dataset. File Length specifies the file size of the dataset in bytes. And, Number of Characters represents the total number of characters in the dataset. The file type of the dataset used is in JSON format and is indicated by File Type Taken. In the literature, there are many studies using dataset in JSON format [3, 14, 15, 19]. File Row Count indicates the size of the number of rows of the dataset used. Number of Words represents the total number of words in the dataset used. Used Software indicates the type of software used to obtain the dataset used. To addition on this, in the 3.2 section of this article, detailed information about the software used is given.

Table 3.1. Properties of the dataset used [51]

No	Explanations	Results
1	Data Capture Start Date	13 April 2017 13:20:43
2	Data Capture End Date	16 April 2017 02:04:11
3	Number of Users	221 756
4	File Length (bytes)	1 019 742 837
5	Number of Characters (excluding spaces)	1 018 855 813
6	Captured Data Size	972 MB
7	File Type Taken	JSON
8	File Row Count	443 512
9	Number of Words	107 307 574
10	Used Software	Python

Thanks to the API [22] created by Twitter, application developers can use Twitter data in their work [23]. Working data clusters were created thanks to Twitter's interface for creating a dataset. Twitter dataset is a huge dataset having the details of 221 756 user accounts. This dataset was extracted from Twitter by using Python software in Json format. Table 3.1 shows the properties of the obtained Twitter dataset. The properties of the dataset are given in Table 3.1. Twitter dataset was opened in Microsoft Excel 2016 [46]. It was randomly selected the tweets by using Excel. And, 81 317 Twitter accounts are randomly selected. They were used in this proposed hybrid spam detection framework model. In this model, short link analysis, machine learning and text analysis methods were used together. Of the remaining 140 439 accounts, random accounts were selected, among which 1 225 Twitter accounts were allocated for use in the training set. There were spam accounts in the dataset obtained from Twitter. A crowdsourcing labeling method was needed to determine spam accounts [8, 47, 48]. The crowdsourcing method was used for checking sensitive content in the dataset and accounts suspended by Twitter. The dataset was obtained from the dynamic structure of Twitter. Spam and not-spam accounts in the dataset obtained from Twitter. It is tried to be determined by comparing the features of real accounts that are labeled as spam by Twitter. The attributes of the user accounts in the training set were extracted. Crowdsourcing method is the gathering of people in a network on the internet to run a specific project [8, 10, 47, -, 50]. In this model, a list containing "screen_name" of the accounts in the dataset was created. Whether these accounts on twitter were suspended or not was checked one by one. Crowdsourcing is a process in which a project is resolved via the internet through participants [10, 48]. They impartially examine the projects and draw a conclusion [8]. The common conclusion reached is used as a labelled part of the project to be implemented. The crowdsourcing method draws a conclusion from the subject analyzed. The conclusion is the common thought of more than one expert. In order to be used in the machine learning method, it is necessary to check the accounts suspended by Twitter. Crowdsourcing method was used to identify the accounts, which are suspended by Twitter and used in the dataset [51]. When a dataset is created via Twitter, some passwords on Twitter API are needed. Consumer Key, Consumer Secret, Access Token and Access Token Secret features show API keys of @oguzhancitlak Twitter account. The API key is different for every Twitter user who opens an account [22, 23, 51]. Twitter uses the json format for processing raw data. Json is built on two structures. This structure is a name / value double collection and an ordered value list. The code in Equation 3.1 is shown to obtain the dataset in json format.

```
def __init__(self, data_dir, query): query_fname = format_filename(query) self.outfile = "%sstream_%s.json" % (data_dir, query_fname) def parse(cls, api, raw):status = cls.first_parse(api, raw)
setattr(status, 'json', json.dumps(raw)) return status
```
(3.1)

The attributes used in a Twitter account are shown in Table 3.2. The method used to determine spam accounts

ERKEN GÖRÜNÜM

is the wait and check method. One year after the dataset was created, in April 2019, the user names of the random accounts selected from the dataset were checked one by one on Twitter. The accounts that were active when the dataset was first created were suspended by Twitter one year later. The attributes of one of the accounts in the dataset extracted from Twitter are used. It is an account suspended by Twitter. The properties of the suspended account are shown in the Table.3.2

Table 3.2. The properties of the suspended account in dataset.

	Features Used	Metric Ranges (Arff File)	Example Account
1	Id	-	"852624385819131904"
2	Screen name	-	"donnelllebla41"
3	Text	-	"#free https://t.co/8WapCjKocF"
4	Name	-	"Alison Scott"
5	User Status Count	0-99,100-199, 200-299, ..., 700-799, …, 1000-1999, 2000-2999, 3000- 3999, …, 5000- 5999, …, 9000-9999, …, 20000-29999, 30000-39999,…, 100000-1999999	"2356"
6	Sensitive Content Alert	TRUE/FALSE	"FALSE"
7	User Favorites Count	0-9,10-19,20-29, 30-39, 40-49, …, 900-999, 1000-1999, 2000-2999, …, 6000-6999, …, 10000- 19999, …, 100000-1999999	"0"
8	User Listed Count	0-9,10-19,20-29, 30-39, …, 80-89, …, 700-799, …, 900-999	"0"
9	Source in Twitter	YES / NO	"twitter.com"
10	User Friends Counts	0-9,10-19,20-29, 30-39, 40-49, …, 100-199, 200-299, …, 400-499, 500-599, …, 1000-9999	"18"
11	User Followers Counts	0-9,10-19,20-29, …, 100-199, 200-299, …, 800-899, …, 900-999, 1000-1999, 2000-2999, …, 10000-19999, …, 30000-39999,…, 100000-1999999	"18"
12	User Created at	2006-2008, 2009-2011, 2012-2014, 2015-2017	"Tuesday August 25 16:36:39 +0000 2011"
13	User Location	YES / NO	"NO"
14	User Geo Enable	TRUE/FALSE	"FALSE"
15	User Default Profile	TRUE/FALSE	"FALSE"
16	Re Tweet	TRUE/FALSE	"FALSE"
17	Suspended Account	TRUE/FALSE	"TRUE"

The attributes of the analyzed dataset were omitted. Table 3.2 shows the required properties of the dataset used in order. Suspended Account was that spam. Twitter suspended spam accounts as they posed security risks to all Twitter users [7]. The attributes determined were used in machine learning method. In the application algorithm of the spam account detection model that has a hybrid structure directed at social networks, the model created has three components. These components; link analysis, machine learning and text analysis. It is shown the framework working diagram in Figure 3.1.

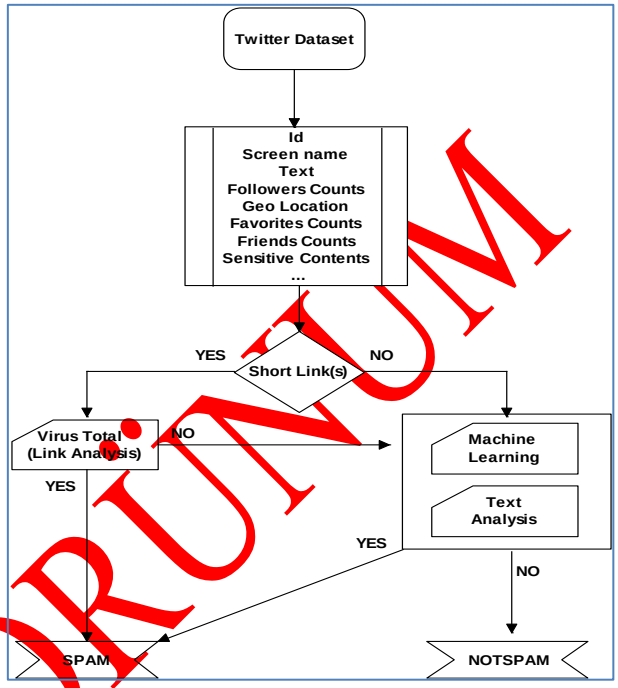

Fig. 3.1. Application working diagram.

First of all, a dataset was created via Twitter for the proposed model [51]. In the steps shown in the Figure 3.1, the attributes of the user accounts in the analyzed dataset are used. The messages in the analyzed accounts are checked whether they contain any short links. Accounts with short links are analyzed in Virus Total [5, 6, 52]. It is the largest online malicious link scanning service [52]. If the account analyzed from the dataset shared a link in message, spam analysis is performed in the Virus Total model. Shared link is analyzed in Virus Total system. If the link analysis model finds malicious sharing in the link, the sharing account will be marked as spam. Otherwise, the link-sharing account is examined in machine learning and text analysis model. Moreover, the accounts that do not contain links in their messages are subjected to the machine learning method. Analyzed account is examined in machine learning model. The account analyzed after the machine learning model is marked as spam. Then, account not marked as spam is analyzed in the text analysis model. The account leaving the machine learning process is evaluated in the text analysis method in order to determine the sensitive contents in the messages. Spam or non- spam Twitter accounts are endeavored to be detected based on the datasets trained on framework. In addition, it is checked whether the account being examined has a sensitive content [53, -, 58]. The dataset processed in the framework is evaluated based on five commonly used

İNCELEN GÖRÜNÜM

algorithms in the literature. These framework algorithms are Naïve Bayes algorithm [34, 59], Random Forest algorithm [60], IBk algorithm [61, 62], J48 decision tree algorithm [63] and JRip algorithm [63, 64]. In the Naïve Bayes classifier algorithm, it is clear which classes the training datasets and other clusters to be classified. Naïve Bayes classification aims to determine the class, or category, of the data presented to the system with a series of calculations defined according to probability principles [59]. The way the algorithm works is it calculates the probability of each state for an element and classifies it according to the highest probability value [34]. It is one of the most restrictive classification techniques available [65]. It is a proven algorithm that is very successful in text classification. Multiple decision trees are used in the Random Forest algorithm [60]. This learning algorithm suggests combining the results of training each of the multivariate decision trees with different sets of training. Another name for the IBk algorithm is (K-nearest neighborhood-KNN) algorithm [61]. It is an algorithm widely used for dataset classification. According to the features extracted during the classification, the new classifier can be explained as the proximity of the K class to the old classifier individuals. The J48 algorithm starts with the existing examples in the dataset. The new decision tree creates the data structure in order to classify new situations. Each node in each row of the decision tree contains a test, which is used to determine which branch to follow after the node. JRip algorithm is one of the most basic and popular algorithms in machine learning [64]. It is an algorithm that creates a set of rules that encompasses all members of the dataset. In this rule determined, it handles the decisions and all the examples of the training data in the dataset as a class. It then proceeds to the next class and performs the same operation here, repeating all until it finds all the classes. Recall, gives information about how much of the information that needs to be brought is brought. Precision is calculated as the ratio of the number of True Positive-TP samples estimated as class 1 to the total number of True Positive-TP and False Positive-FP to all sample numbers estimated to be class 1. Accuracy-Error Rate has an important place in measuring the performance of the model used. This error rate gives the model's accuracy rate as the simplest and most popular of the dataset processed. F-Measure is the harmony mean of the Precision and Sensitivity values. These criteria alone are not sufficient to draw a meaningful comparison result. Kappa statistics is a statistical method that measures the reliability of agreement between two values. The number of k varies between -1 and +1. Full compliance in the datasets it is used occurs when K is equal to 1. The framework has three components.

3.1.2. Short Link Anaysis Method

Virus Total is able to test the connectins directed to it thanks to its security datasets it contains within. It scans the dataset of more than sixty security companies at the same time [52]. It basically tries to detect whether a URL link is malicious or not by scanning it with antiviruses. It performs malicious link detection by comparing it with the signatures in its database [5, 6, 31, 33]. It has a dynamic structure. It constantly updates itself thanks to its dynamic structure. For example, a malicious website released today may not be immediately detected by the services that support the database of Virus Total [22]. Google Safebrowsing [66], Kaspersky, OpenPhish, ComodoSite Inspector, Forcepoint ThreatSeeker, Opera and Yandex Safebrowsing [66] services are Virus Total's powerful malware link analysis services [22, 67]. Unfortunately, a new malicious link currently shared via social media will not be detected unless it has been defined in the Virus Total dynamic structure. However, Virus Total is widely available in the literature in malicious link analysis system [5, 6, 32, 68, 69].

```
import requests
url = 'https://www.virustotal.com/vtapi/v2/url/scan'
params = {'apikey': '2730d9196c7a8fad7cbdeeead2[redacted]59924919b056a3481314325 ', 'url':'<url>'}
response = requests.post(url, data=params)
print(response.json())
```

Fig. 3.2. This is the API code used in the virus total [52, 70].

API code is important in short link analysis. Figure 3.2 shows the API code used in python [70]. Actually, all the codes used are ready and provided by Virus Total [52]. In the short link analysis method, the link in the messages sent from the user account is important. If there is a short link in the transmitted text, this account is sent for short link analysis [7]. The malicious link contained in a message is enough to mark this account as spam [3, 7, 25, 32, 67]. If there is no link in the message of the analysed account or if the link analysed by Virus Total is not specified as malicious, the proposed model sends the analysed account to the framework repository where the machine learning and text analysis stages are applied. In Figure 3.3, a social media account is explored on the application. A software was developed for the created model.

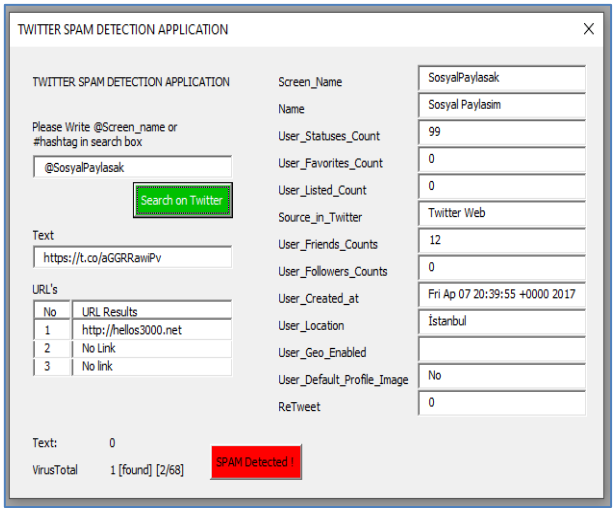

Fig. 3.3. A malicious link detection screen on the application.

ERKEN GÖRÜNÜM

It is understood that the social media account analyzed in the Figure 3.3 shares a malicious link in its message. It is the application that detected a malicious link in the URL section. The URL in the analyzed Twitter account was sent for scanning using the API Key provided by Virus Total. Values returned from Virus Total API [70] are in JSON format. The Virus Total system detected this malicious link shared in the message as malicious [7, 67]. There is a malicious link in the message. Link analysis method had been very successful, as all 1 870 accounts with malicious links found were also suspended by Twitter. It is that the Virus Total dataset is dynamically structured. There is an important situation that even if a real user deliberately shares a malicious link, this account may be considered spam according to the spam policy of Twitter [7]. Submitting a URL with the Virus Total API does not always produce a result. In this model, there could not be get a response from only 31 accounts. Figure 3.1 shows the diagram of the proposed framework. If no results can be obtained from URL analysis, the framework continues to work with machine learning and text analysis methods.

3.1.3. Machine Learning Method

Machine learning is capable of interpreting very large datasets with a large number of attributes. The machine learning method performs this work when there are no suitable equations and functions to interpret large datasets [11, 21, 45, 51, 71, 72].

Some parts of the values in the dataset with Arff extension used are shown in Table 3.3. Metric ranges of the attributes used in the dataset are shown in the Figure 3.3. The total number of tweets sent by the analyzed accounts is indicated by User_Status_Count. For example, this attribute is in the first line 700-799. This shows that the number of tweets is between 700 and 799. In the method purposed, this metric range shows the total number of Tweets in the simple account messages used. In another example, the range of the "user_listed_count" attribute is given as 0-9. This range shows the list of contacts owned by the simple account analysed. The Arff file whose properties are shown in Table 3.3 also has 23 650 words. This file is used in machine learning model. It has a total of 1 276 lines and has 129 132 characters after the spaces were omitted. The size of the file is 130 410 bytes. The name of the attributes used in the Table 3.3 are same as the name of the metrics in the Twitter dataset. The features in the dataset with Arff extension in Table 3.3 are as follows. @relation is on the first line of the file and indicates the file type. @attribute shows the properties of the dataset starting from the second line. @data shows all details in file. There is a link in T.CO format in the tweet sent by the simple account. The Virus Total system analyzed whether or not this account was malicious but was subjected to machine learning because the link in the message is unsuspecting. Attributes of the Twitter user of the simple account were evaluated in machine learning method. The dataset was split into two parts for training and testing. There was a similar situation in the studies in the literature. For machine learning models, the dataset must be split. It was divided

Table 3.3. Arff details of spam and non-spam accounts analyzed as a sample.

No	User Status Count	Sensitive Content Alert	User Favorites Count	User Listed Count	Source in Twitter	User Friends Counts	User Followers Counts	User Created at	User Location	User Geo Enable	User Default Profile Image	ReTweet	Suspended Account
1	30000-39999	FALSE	900-999	30-39	NO	1000-1999	1000-1999	2012-2014	NO	FALSE	TRUE	FALSE	FALSE
2	2000-2999	FALSE	0-9	0-9	YES	10-19	10-19	2009-2011	NO	FALSE	TRUE	FALSE	TRUE
3	5000-5999	FALSE	6000-6999	0-9	YES	400-499	900-999	2012-2014	NO	TRUE	FALSE	TRUE	FALSE
4	3000-3999	FALSE	40-49	0-9	YES	0-9	20-29	2009-2011	NO	FALSE	TRUE	FALSE	TRUE
5	1000-1999	FALSE	10-19	0-9	YES	0-9	20-29	2012-2014	NO	FALSE	TRUE	FALSE	TRUE
6	2000-2999	FALSE	0-9	0-9	YES	0-9	10-19	2009-2011	NO	FALSE	TRUE	FALSE	FALSE
7	1000-1999	FALSE	10-19	0-9	YES	0-9	0-9	2012-2014	NO	FALSE	TRUE	FALSE	TRUE
8	30000-39999	FALSE	10000-19999	0-9	YES	500-599	1000-1999	2012-2014	NO	TRUE	TRUE	TRUE	FALSE
9	2000-2999	FALSE	30-39	0-9	YES	200-299	200-299	2015-2017	NO	FALSE	TRUE	FALSE	TRUE
10	20000-29999	FALSE	2000-2999	80-89	YES	3000-3999	2000-2999	2012-2014	YES	TRUE	TRUE	TRUE	FALSE

ERKEN GÖRÜNÜM

to 66% training and 33% testing. This rate, were preferred, was generally accepted in the literature [71]. According to the crowdsourcing model, 1 225 Twitter accounts were labeled as spam. The features of these spam accounts were used in the training set of machine learning. A few of the suspended accounts were as follows @LeeahWStudlos @Brettacus268 @footystory @footywnews @premnewuk @tmllesmarker @fbtips @priceboostbets.

In Weka, which is one of the machine learning methods, the value of the K-Layered Cross Validation option is taken as 10 in many studies [71, -, 75]. In order to evaluate the success of the method applied in the studies conducted in Weka, the dataset is divided into two as training and test sets. In order to test the success of the dataset, 66% training, 33% partitioning as test set, testing the success with the test set after the system is trained can be used as another method. K-fold cross validation is a method used to evaluate the success of machine learning models. Before using this method, the K value is set, this value is usually given as 10 by default. In our literature researches, it is very common to use the K value as 10 based on the K-Layer Cross Validation method [72, 74]. Our dataset for K 10 is primarily divided into 10 equal parts. And 10 results were obtained from each part separately. The validated data was used at the same time with the test data each time in parts. After the entire dataset is divided into parts equal to our K value, the K-Layer Cross Validation system starts to work. First of all, we want to train our model with a part of dataset and evaluate the success of our model with a part of it. One of our 10 pieces was randomly chosen and the rest was used for training. Depending on the dataset, some biases and errors might occur in the validation and testing of the model. But, it doesn't matter what part you start from. K-fold cross validation splits data into equal parts based on a specified “k” number, it allows each part to be used for both validation and testing. As a result, the same method was applied 10 times in 10 different validation and test sets. The performance or overall error rate of the proposed spam detection model is calculated as the average of these 10 results. In the K-Layer Cross Validation method, the formula in Equation 3.2 is used in calculating the results of S1, S2,…, S10.

$$t_i \in VK \text{ to be}, \quad Result = \frac{\sum_{i=0}^{k} SF\,(ti,\ VK - ti)}{k} \qquad (3.2)$$

In this formula, each test set selected from the dataset "t", "k" shows that how many pieces of folding are used, "VK" indicates the dataset and "SF" classification function. K-fold cross-validation was used with both a validation and test set. Validation dataset was shown as VK and “t” was referred to validation plus test dataset. “i” was data at time “t”. The total data set is split into k sets in eqn 3.2. In Weka, Preprocess, Classify, Cluster, Associate, Select Attributes, Visualize panels are used. The dataset has 1 225 Instances and 13 Attributes. The parameters and scaling range of the training set used are given in detail in Table 3.2. In addition to this, in the random sampling method, it was also known as Monte Carlo Cross Validation (MCCV) [76]. the data set is divided into training and test data sets at rates determined by the user k times [76, 77]. The difference from cross validation is that k random but specific partitions can partially or completely correspond to the same points in the dataset. MCCV does not work properly in Weka [76, 78]. The cross validation model was chosen according to the studies in the literature [72, 74].

3.1.4. Text Analysis Method

In text analysis, it is aimed to reach statistical results through the text. This method aims to obtain the data from Twitter's dynamic dataset. The sensitive content word "bet" is used in the dataset collected from the Twitter repository with the Python tool. It aims for studies such as asset relationship modelling, text summarizing, emotion analysis, extracting topics from the text, the production of class particles, classification and segmentation [21, 26, 79, 80]. In the text analysis method, the messages of the accounts analysed are taken into consideration. In the literature, words of spam accounts frequently used are determined [53, -, 58]. Sensitive words out of these words are determined by their frequency of being mentioned in the messages. These words were used in the text analysis method. When the “#” symbol was placed at the beginning of any word on Twitter, the tweet was categorized with the word with the "#" symbol. Categorized words could be listed much easier on Twitter. Sensitive designed words were chosen from the words frequently used in the dataset. The designated words are listed below.

“#free, #mature, #porn, #hot, #events, #sex, #nude, #adult, #teen, #pornvideos, #erotic, #betta, #Bitcoin, #pussy, #ass, #windows, #boobs, #hooker, #tits, #massive, #um, #babe, #beat, #casino #bonus #naked, #gays, #eroticism, #fantasy, #panty, #tips, #blockchain”

An account with spam words in almost every tweet can be considered to be a spam account. Text mining is used to classify twitter account as spam and non-spam. In order to realize the classification process, it is necessary to prepare the text with text mining. Tokenization, Lemmatization, Term weighting and Feature selection preprocesses are used in text analysis method. Tokenization is a process used to extract words from a twitter text [75]. Lemmatization is a process for grouping and calculating identical words [75]. Term weighting is the frequency at which an attribute is observed in a document [81]. In the text analysis method, it is examined whether the previously identified sensitive words are mentioned in the messages. Figure 3.4 shows a social account that shares sensitive content in its message.

TRENDOKUM

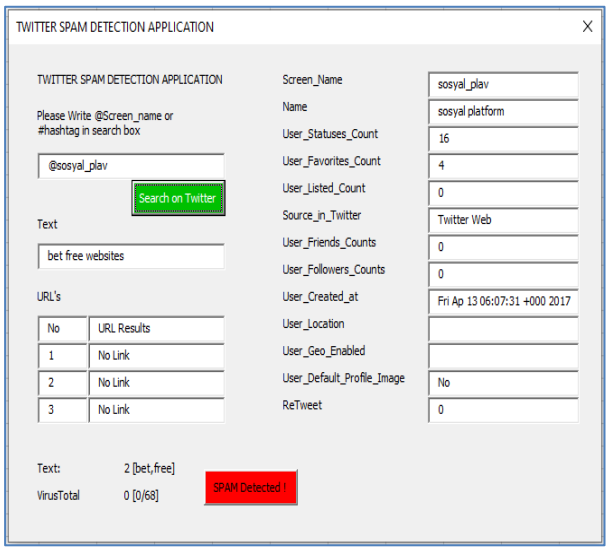

Fig. 3.4. An account that shares sensitive content on the application.

In the Figure 3.4, a message containing two sensitive words is shared by the account analyzed. A message containing bet and free was shared by the account @sosyal_plav. There is no link in the message of the said account. It passed thanks to the link analysis method as an unsuspecting user because there was no link in the message. The accounts with sensitive content in the Twitter dataset were thus determined. If the machine learning process is performed based on the attributes shown in the account details section, it turns out that this account is a spam. In addition to this, the contribution index was used to express the imbalance in the messages sent on Twitter in the text analysis part of proposed framework [82, 83]. It measured to identify spammers by comparing to spam messages to each other [82]. This value varies between [1, −1]. 1 contribution index in this range identifies only people who posted a tweet or retweet without receiving a reply. In the text analysis part, it was not matter whether the analyzed accounts responded back to their tweets or not [83]. The formula of contribution index was shown in equation (3.3).

Contribution Index (Ci) to be,

$$Ci = \frac{Tweets_{sent} - Tweets_{received}}{Tweets_{sent} + Tweets_{received}} \quad (3.3)$$

This measure was meant to express the unbalance in messages sent when compared to messages received by spammers. A high contribution index should be used to usually identify spammers. The contribution of the node "Ci" is defined by the formula in eqn 3.3 In this study, even if the selection of the sensitive tweets is carried out by crowdsourcing method, the authors often noticed that accounts with a contribution index greater than 0.76, posted a significant number of sensitive tweets.

4. RESULTS AND DISCUSSION

In the dataset in the machine learning method, the attributes of 1 225 Twitter user accounts labelled based on the Crowdsourcing method are defined. This dataset in the machine learning process is evaluated in two different ways with K=10 and percentage split = 66, it is very common to use the K value as 10 based on the K-Layer Cross Validation method that means quite acceptable in framework according to literature [72, 74]. K and percentage split values are used as training set and test set in framework. The proposed model uses five commonly used machine learning algorithms in the literature. These are Naïve Bayes, Random Forest, IBk, J48 and JRip. The number of accounts that do not share links is 11 385. The number of accounts sharing unsuspecting links is 68 062. A total of 79 447 accounts were transferred for Machine learning and Text analysis methods based on the application algorithm of the proposed model. Since all of the 1 870 malicious accounts that Virus Total found are also suspended by Twitter. The results of the algorithm and measurement metrics we achieved as a result of the attributes of the simple account, which is @FOXHaber, analyzed in the machine learning model are shown in Table 4.1.

Table 4.1. Analysis of a social media account through machine learning.

Feature	NaiveBayes	IBk	J48	Result
Correctly Classified Instance	91.9951 %	99.0295 %	97.9863 %	96.3369 %
Incorrectly Classified Instance	8.0049 %	0.9705 %	2.0137 %	3.6631 %
Kappa Statistic	0,8373	0,98	0,9581	0,9251
Accuracy-Error Rate	0,4408	0,0001	0,0241	0,155
Precision	0,923	0,990	0,973	0,962
Recall	0,918	0,990	0,971	0,959
F-Measure	0,919	0,990	0,971	0,960

Account	Spam	Not Spam
A simple account	3.6631 %	96.3369 %
Result	Not Spam	

The social account, which is analyzed in the Table 4.1 and whose attributes are given in Figure 3.4, is considered to be non-spam as a result of the analyses. As a result, it had been shown the probability of this account being spam as approximately 3% according to proposed model. In other words, it had been seen that this account was approximately 97% not-spam. The analysis was performed on Twitter dataset through IBk algorithm (99.0385%), which yielded the best success rate previously, Naïve Bayes algorithm yielding the lowest success rate (91.8269%) and J48 algorithms (97.1154%) that yielded a mid-range success rate. According to the results examined in Table 4.1, it is determined based on the results of these three algorithms that the account is not a spam account. For these three algorithms, Kappa

ERKEN GÖRÜNÜM

statistic yields a 0.9251 ratio. This spam analysis process has a value very close to 1. The accuracy error rate of 0,155 appears quite low and is valid for successful results. It can be decided by analyzing these metric results that the analyzed account is not spam. When the Kappa statistic is close to 1, it is seen that the correct result is reached [48]. The ratio of correctly predicted to total correct and incorrect predictions TP / (FP + TP). Precision shows how many of the values predicted as Positive are actually Positive. The average of the Precision value of the account analyzed in Table 4.1 is 0,962. As Recall, it is a metric that shows how much of it should be predicted as Positive. The average of the Recall value of the account analyzed in Table 4.1 is 0,959. F-Measure is the average Harmony of Precision and Recall values. The average of the F-measure value of the account analyzed in Table 4.1 is 0,960. When the correctly and incorrectly classified metrics are evaluated, it is revealed that the analyzed account is not a spam account by 96,3369%. The mean absolute error rates are far from 1. This value is very close to 0 and is very useful in interpreting the result obtained in machine learning operations. In the analysis of the proposed model in machine learning is made and the whole of the dataset in hand and the measurement results of other metrics with respect to split = 66 and K = 10 are shown in Table 4.2.

Table 4.2. Comparison of Dataset with respect to percentage split = 66 and K = 10.

Metrics	*NaiveBayes (percentage split = 66)*	*NaiveBayes (K = 10)*	*Random Forest (percentage split = 66)*	*Random Forest (K = 10)*	*IBk (percentage split = 66)*	*IBk(K = 10)*	*J48 (percentage split = 66)*	*J48 (K = 10)*	*JRip (percentage split = 66)*	*JRip (K = 10)*
Correctly Classified Instance	*91.8269%*	*92.1639%*	*97.8365%*	*99.1837%*	*99.0385%*	*99.0204%*	*97.1154%*	*98.8571%*	*96.875%*	*98.6939%*
Incorrectly Classified Instance	*8.1731%*	*7.8367%*	*2.1635%*	*0.8163%*	*0.9615%*	*0.9796%*	*2.8846%*	*1.1429%*	*3.125%*	*1.3061%*
Kappa statistic	*0,8341*	*0,8405*	*0,9556*	*0,9831*	*0,9802*	*0,9798*	*0,9399*	*0,9763*	*0,9355*	*0,973*
Accuracy-Error Rate	*0,1062*	*0,094*	*0,0325*	*0,0171*	*0,0091*	*0,0089*	*0,0394*	*0,0223*	*0,0391*	*0,019*
F-Measure	*0,919*	*0,922*	*0,978*	*0,992*	*0,990*	*0,990*	*0,971*	*0,989*	*0,969*	*0,987*

In the Table 4.2, when the percentage split is taken equal to 66 in the algorithms used in machine learning method, it is seen that IBk algorithm yields the most successful results. This assumption is arrived at for the fact that the percentages of kappa statistic, accurate approved rates and accurate classification are higher than the other four algorithms. The fact that the mean absolute error rate is negligible is another supporting data. In this study, it is seen that Naïve Bayes classification algorithm yields the lowest result among the other four algorithms. When K=10 is taken in the algorithms used in machine learning, looking at Correctly Classified Instance, Incorrectly Classified Instance, Kappa statistic, Accuracy- Error Rate and F-Measure values in Table 4.2, it is seen that Random Forest is a more successful algorithm than others. This result can be understood by looking at these five metrics in the Table 4.2. Naïve Bayes classification algorithm yields the lowest result among the other four algorithms. The algorithms mentioned in the Table 4.2 are used in the analysis of accounts in the same Twitter dataset. In each method of the proposed model, the same dataset is processed, therefore K=10 and percentage split=66 values are obtained close to each other in the methods. The running logic of each algorithm is also different. For example, Random Forest and J48 are decision tree algorithms. When a comparison is made among the five algorithms used, it is understood that the IBk algorithm is the most successful algorithm with 99.0295 % average results. Nonetheless, the Naïve Bayes algorithm has the lowest average of 91.9951 % among the five algorithms used. These five algorithms evaluated yield a success rate of over 90% according to results we have. In machine learning applications, values of 90% and above are considered quite successful [11, 72]. Text analysis made on Twitter accounts, which are detected to be spam as a result of the machine learning process is run in a supporting manner [73, 78]. Machine learning and Text analysis are the methods used together and can be evaluated in the same repository. The account subjected to machine learning process can also be processed in text analysis method. 62 073 of 63 895 accounts analyzed in total were marked as spam based on the sensitive content they contained. In this case, the success rate of our text analysis method is 97.15 %. All accounts in the Twitter dataset analyzed in this study contain sensitive content, and all accounts marked as spam meet at least one text analysis condition [26, 36, 50, 75, 79, 80, 84]. For this reason, the aim of our text analysis method is considered to increase the success rate of machine learning. The presence of sensitive content in all accounts to be considered spam shows us that we achieved the same success rate in the text analysis as the success rate of machine learning. In machine learning, the sensitive content within the messages of the specified spam accounts is checked. The algorithms used to measure the performance of the machine learning method and the metric values obtained are shown in Table 4.2. As a result of the evaluations in this Table 4.2, machine learning

ERKEN GÖRÜNÜM

performance is considered to be 97.06%. For the success of the proposed model, the spam account rates in the accounts analyzed by each method are taken into account. The high performance of the algorithms used in the study of machine learning is an indication that the detected calculations are correct with little negligible errors. Table 4.2 shows the learning algorithm results we obtained. As a result of the evaluation of link analysis, machine learning and text analysis methods, the success of a spam account detection model based on the social networks that we developed is 95.69%. Table 4.3 shows the success details of the proposed model.

Table 4.3. Component details of the model proposed.

No	Methods	Spam	Not Spam	Totals	Method Success Rate	Model Success Rate
1	Link Analysis	1 870	68 062	69 932	98.37%	95.69%
2	Machine Learning	14 237	1 315	15 552	91.54%	
3	Text Analysis	62 073	1 822	63 895	97.15%	

The proposed model in Table 4.3 has three components. In the short link analysis method, all malicious links were detected by Virus Total. In this analysis model, there could not be get a response from only 31 accounts. It was added these accounts in calculation. Therefore, the success rate of method approximately was 98.37%. Virus Total is used in malicious link analysis. It has a dynamic structure. These 31 accounts, which are known to be really spam, could not be detected. This has affected the success rate. The machine learning model of the proposed hybrid model is more successful than the Wang [15] study despite using similar algorithms. The success of the machine learning after the controls appears to be 91.54%. The success of the text analysis is calculated as 97.15%. When the total success of the model is taken as the average of these three components, it is calculated as 95.69%. The dynamic structure of link analysis can change the model success rate. All of accounts marked as spam in link analysis shared malicious links. The remaining 68 062 accounts still contain spam, but link analysis cannot detect them. Malicious links have to identify this by machine learning and text analysis methods. In the remaining 68 062 accounts, there are no malicious links according to Virus Total. All of the malicious and unsuspecting accounts were detected owing to Link analysis method. As a result of machine learning, the algorithm with the lowest success rate exhibits 91.99% performance. In the calculation of model performance, the average of the performances of three models is taken. The success rate of the short link analysis component is 100%, the success rate of the machine learning component is 91.54% and the success rate of the text analysis component is calculated as 97.15%. The success rate of the proposed model is 95.69 % based on these results.

The comparison of the spam detection methods in the literature and the studies carried out on this subject based on the research carried out in this study is shown in Table 4.4. In these studies, the ways in which spammers threaten the users' personal data and the methods by which they perform this are examined in detail and the results are presented in Table 4.4. When the Table 4.4 is examined, it is understood that different methods are used in the detection of spam accounts in social networks based on the comparisons made. The most recent dataset is used in the proposed model [51]. In addition to this, accuracy rates in similar datasets are shown It is aimed to show a general summary of the studies done in the literature part of the manuscript with the proposed model. Most of the studies used methods based on machine learning.

Table 4.4. Comparison of spam detection studies in the literature.

Article	Technical	Algorithm	Evaluation Metric	Used Dataset	Accuracy Rates
Akiyama and ark.[13]	Monitoring System	Web-based Genetic Algorithm	Performance Rate	URL Dataset in which Routing Codes are Injected-Malicia (2013)	95.50%
Wang [15]	Machine Learning System	Naive Bayes Algorithm and Twitter API	F- measure	Trend Topics in Twitter dataset (2010)	89.00%
Romo and Araujo [16]	Machine Learning System	Traditional Classification Algorithm	F- measure	Trend Topics in Twitter dataset (2012)	94.50%
Liu and ark. [17]	Machine Learning System	Graph-based Algorithm	True Positive	Twitter and URL Datasets (2016)	78.00%
Lee and Kim [18]	Machine Learning System	Creating and Supporting Vektor Machine Algorithm	False Negative	Twitter dataset (2011)	86.53%
Miller and ark. [19]	Machine Learning System	DenStream and StreamKM++	F- measure and Precision	Labeled Twitter dataset (2009)	95.55%
Maio et al. [20]	Deep Learning, CmpNN, Word2vec	Twitter API, Comparative Neural Network	MAP, NDCG	(5 000 tweets) (26/01/2017 - 12/02/2017) Twitter Dataset	93.8%
Chatterjee et al. [21]	Deep Learning, CNN, LSTM, Glove, Word2Vec, FastText, Sentiment Specific Word Embedding (SSWE),Microsoft Cognitive Toolkit	Semantic word embedding, Sentiment word embedding, Naive Bayes (NB), Gradient Boosting Decision Tree (GBDT), Support Vector Machines (SVM),	F-Measure, Recall, Precision,	Twitter Dataset (17.62 Million tweet 2012-2015)	93.2%
Proposed Model	Machine Learning, Link and Text Analysis	Machine Learning Algorithms and Twitter API, Virus Total	F- measure and Precision, Sensitive content rate,	Twitter and URL Datasets (2017)	95.69%

ERKEN GÖRÜNÜM

The machine learning method is an important component of our spam account detection model in the social networks we recommend. The study conducted by Miller et al [19], in 2014 shows a high performance. The most significant and important feature of this study is that combination of two classification algorithms were used. Akiyama et al [13] use a different system and algorithms in their study. They get very good results on the dataset they used. In fact, this study shows us that higher results can be achieved with more datasets and changing algorithms in the years to come. Social networks are used by millions of users. The datasets obtained from these networks can reach very large sizes. When we look at Table 4.4, systems built on machine learning can analyze these large data sets. In the methods where the machine learning methods are used, similar evaluation metrics are used so that we can make a more accurate comparison between studies. Again, in the studies examined, real Twitter datasets, URL information, and profile information on social media are also used for spam detection. It is seen, among these data, that studies using real Twitter datasets have higher performance rates. In the studies carried out by the classification method, it is seen that accounts with spam messages and some messages belonging to real users who do not normally produce spam are considered as spam. The performance of the proposed hybrid spam detection framework in this study is calculated to be 95.69 %. This ratio is in the top of the studies examined in Table 4.4. It can be said that this is due to the use of multiple evaluation metrics. Existing studies used only one technique, but in the proposed model, three different techniques were combined to process a new model. For this reason, it can be stated that it shows a higher performance rate than the current studies. The success rate of Miller's work is 95.55 %. The success rate of the proposed hybrid model is 95.69 %. These two works are very close to each other. However, the success of the proposed model is higher. Accounts in the dataset used in the proposed model were obtained online from Twitter and have spam accounts that cannot be detected instantly by Twitter. In the evaluations performed, these possible spam accounts were ignored. However, it is considered that the performance rate of 95.69 % achieved when this dataset was analyzed will remain below the performance rate to be obtained by re-analysis of the same dataset at later times. The success rate of Miller et al's study [19] and the proposed model are very close to each other. They used a very old dataset in their study. Moreover, it is aimed to detect spam and malicious accounts in all of the studies in Table 4.4. In this study, more than one model is proposed together. Each model has its own limitations. However, in general, the limitation of the proposed framework model is in the link analysis model. In this model, the API used has some limitations. These limitations are set by Virus Total [52]. Ability to send 500 requests per day and 4 requests are processed per minute [52, 67, 70]. When analyzing a large dataset, it may take a little more time due to these limitations.

5. CONCLUSION

In this study, the performance of the proposed learning-based spam account detection framework on the dataset and live social networks is examined. The accounts marked as a result of each method were again questioned on the social network, taking into account the difference between the time of creation and analysis of the dataset. Over two hundred thousand user accounts were obtained from Twitter. In the developed framework, measurement metrics, advantages and disadvantages, algorithm, technical details and the accuracy values obtained were compared. The results of the methods used in the literature have been analyzed. By contrast, we found that our successful results were higher than others result. The number of malicious and non-malicious links is specified in the short link analysis section. 1 870 malicious links were obtained. The average performance of link analysis, machine learning and text analysis methods is considered as the success rate of the framework. In proposed framework, three different components that we use have checked spam account in three different ways. Therefore, it is important to use three different models together. It is aimed to increase the performance of the proposed model by using three different models together. In the light of all the analyses carried out and the results, it can be calculated that the performance rate of the proposed model is 95.69%. Furthermore, it is considered that analysis of previously extracted datasets is as important as the analysis of live social network datasets in order to develop a social network spam detection policy. In many studies in the literature, spam detection on social networks can be done on datasets. However, in future studies, spam detection in social networks will be possible immediately without a dataset. Thus, spam accounts can be detected in social platforms with a very short time.

DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Oğuzhan ÇITLAK: Obtained the dataset on the social network, performed the models, analysed the results and wrote the manuscript.

Murat DÖRTERLER: Determined the algorithm used in the study, checked the models used and observed the result obtained.

İbrahim ALPER DOĞRU Checked the studies in the literature the manuscript, determined the algorithms used and played a role in the creation of the model.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

ERKEN GÖRÜNÜM

REFERENCES

[1] Erdoğan G. and Bahtiyar Ş., "Sosyal ağlarda güvenlik", ***Akademik Bilişim Konferansı,*** 1-6, (2014).

[2] https://makeawebsitehub.com/social-media-sites/, "95+ Social Networking Sites You Need To Know About In 2021", (16 January 2021).

[3] Kabakus A. T. and Kara R., "A survey of spam detection methods on twitter", ***International Journal of Advanced Computer Science and Applications***, 8(3): 29-38, (2017).

[4] https://dijilopedi.com/2020-turkiye-internet-kullanimi-ve-sosyal-medya-istatistikleri/, "2020 Türkiye İnternet Kullanımı ve Sosyal Medya İstatistikleri",(17 April 2021).

[5] Wang S., Chen Z., Yan Q., Ji K., Peng L., Yang B. and Conti M., "Deep and broad URL feature mining for android malware detection", ***Information Sciences***, 513: 600-613, (2020).

[6] Hong J., Kim T., Liu J., Park N. and Kim S. W., "Phishing url detection with lexical features and blacklisted domains", ***In Adaptive Autonomous Secure Cyber Systems,*** Springer, Cham, 253-267, (2020).

[7] https://help.twitter.com/en/safety-and-security/phishing-spam-and- malware-links, "About unsafe links Twitter spam or malware links and blocking links", (11 May 2021).

[8] Buecheler T., Sieg J. H., Füchslin R. M. and Pfeifer R., "Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework" ***In The12th International Conference on the Synthesis and Simulation of Living Systems***, Odense, Denmark, MIT Press, 679-686, (2010).

[9] Dent K., and Paul S., "Through the twitter glass: Detecting questions in micro-text", ***arXiv preprint arXiv:2006.07732***, (2020).

[10] Hendal B., "Hashtags as Crowdsourcing: A Case Study of Arabic Hashtags on Twitter", ***Social Networking***, 8(4): 158-173, (2019).

[11] Suzuki Y., "Filtering Method for Twitter Streaming Data Using Human-in-the-Loop Machine Learning",***Journal of Information Processing***, 27: 404-410, (2019).

[12] Mata S. J. I., "Anomaly Detection as a Method for Uncovering Twitter Bots", (2019).

[13] Akiyama M., Yagi T., Mori T. and Kadobayashi Y. "Analyzing the ecosystem of malicious URL redirection through longitudinal observation from honeypots", ***Computers & Security***, 69: 155-173, (2017).

[14] Fernandes M. A., Patel P. and Marwala T., "Automated detection of human users in Twitter" ***Procedia Computer Science***, 53: 224-231, (2015).

[15] Wang A. H., "Don't follow me: Spam detection in twitter", ***In Security and cryptography (SECRYPT), proceedings of the 2010 international conference on***, IEEE, 1-10, (2010).

[16] Romo J. and Araujo L., "Detecting malicious tweets in trending topics using a statistical analysis of language", ***Expert Systems with Applications***, 8: 2992-3000, (2013).

[17] Liu S., Zhang J., Wang Y. and Xiang Y., "Fuzzy-based feature and instance recovery". ***In Asian Conference on Intelligent Information and Database Systems,*** Berlin, Heidelberg, 605-615, (2016).

[18] Lee S. and Kim J., "Early filtering of ephemeral malicious accounts on Twitter", ***Computer Communications***, 54: 48-57, (2014).

[19] Miller Z., Dickinson B., Deitrick W., Hu W. and Wang A. H., "Twitter spammer detection using data stream clustering", ***Information Sciences***, 260: 64-73, (2014).

[20] Demaio C., Fenza G., Gallo M., Loia V. and Parente M.,"Time-aware adaptive tweets ranking through deep learning", ***Future Generation Computer Systems***, 93: 924-932, (2019).

[21] Chatterjee A., Gupta U., Chinnakotla M. K., Srikanth R., Galley M., and Agrawal P., "Understanding emotions in text using deep learning and big data", ***Computers in Human Behavior***, 93: 309-317, (2019).

[22] https://apps.twitter.com/app/13644526/keys, "Twitter API page", (14 May 2021).

[23] https://developer.twitter.com/en/community#, "Twitter Community Developer", (16 May 2021).

[24] Ahmed F. and Abulaish M., "A generic statistical approach for spam detection in Online Social Networks", ***Computer Communications***, 36: 1120-1129, (2013).

[25] Çıtlak O., Dörterler M. and Doğru, İ. A., "A survey on detecting spam accounts on Twitter network", ***Social Network Analysis and Mining***, 9: 1-13, (2019).

[26] Lüdering J. and Tillmann P., "Monetary policy on twitter and asset prices: Evidence from computational text analysis", ***The North American Journal of Economics and Finance***, 51: 100875, (2020).

[27] Karamollaoğlu H., Doğru İ. A. and Utku A., "Identification of shares containing offensive charge in social media", ***In 2017 25th Signal Processing and Communications Applications Conference (SIU),*** IEEE, 1-4, (2017).

[28] Grandjean M., "A social network analysis of Twitter: Mapping the digital humanities community", ***Cogent Arts & Humanities***, 3.1, 1171458, (2016).

[29] Alom Z., Carminati B. and Ferrari E., "A deep learning model for Twitter spam detection", ***Online Social Networks and Media***, 18: 100079, (2020).

[30] Arici N. and Yildiz E., "Gercek Zamanli Bir Saldiri Tespit Sistemi Tasarimi Ve Gerceklestirme", ***Engineering Sciences***, 5.2: 143-159,(2010).

[31] Gupta N., Aggarwal A. and Kumaraguru P., "bit. ly/malicious: Deep dive into short url based e-crime detection", ***APWG Symposium on Electronic Crime Research (eCrime)***, IEEE, 14-24, (2014).

[32] Çıtlak O., Doğru İ. A. and Dörterler M., "A Spam Detection System with Short Link Analysis", ***10th International Conference on Information Security and Cryptology (ISCTURKEY 2017),*** Ankara, 178-185,(2017).

[33] Nepali R. K. and Wang Y., "You look suspicious!!: Leveraging visible attributes to classify malicious short urls on twitter", ***49th Hawaii International Conference on System Sciences (HICSS),*** IEEE, 2648-2655, (2016).

[34] Ren J., Lee S. D., Chen X., Kao B., Cheng R. and Cheung D., "Naive bayes classification of uncertain data. In Data Mining", ***9th IEEE International Conference***, IEEE, 944-949, (2009).

[35]. Simsek M., Yilmaz O., Kahriman A. H. and Sabah L., "Detecting Fake Twitter Accounts with using Artificial

ERKEN GÖRÜNÜM

Neural Networks", ***Artificial Intelligence Studies***, 1.1: 26-29, (2018).

[36] Liu S., Wang Y., Zhang J., Chen C. and Xiang Y., "Addressing the class imbalance problem in twitter spam detection using ensemble learning", ***Computers & Security***, 69: 35-49, (2017).

[37] Kabakus A. T. And Kara R., "TwitterSentiDetector: a domain-independent Twitter sentiment analyser", ***INFOR: Information Systems and Operational Research***, 56.2: 137-162, (2018).

[38] Wu T., Liu S., Zhang J. and Xiang Y., "Twitter spam detection based on deep learning", ***In Proceedings of the australasian computer science week multiconference***, 1-8, (2017).

[39] Henderson P. and Ferrari V., "End-to-end training of object class detectors for mean average precision", ***In Asian Conference on Computer Vision***, Springer, Cham, 198-213, (2016).

[40] Sharma A., Tian Y., Sulistya A., Lo D. and Yamashita A. F., "Harnessing Twitter to support serendipitous learning of developers", ***In 2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)***, IEEE, 387-391, (2017).

[41] Goldberg Y. and Levy O., "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method", ***arXiv preprint arXiv:1402.3722***, (2014).

[42] Arslan R. S. and Barışçı N., "Development of output correction methodology for long short term memory-based speech recognition", ***Sustainability***, 11.15: 4250, (2019).

[43] Pennington J., Socher R. and Manning C. D., "Glove: Global vectors for word representation", ***In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)***,1532-1543, (2014).

[44] Athiwaratkun B., Wilson A. G. and Anandkumar A., "Probabilistic fasttext for multi-sense word embeddings", ***arXiv preprint arXiv:1806.02901***, (2018).

[45] Etaati L., "Deep Learning Tools with Cognitive Toolkit (CNTK)", ***In Machine Learning with Microsoft Technologies***, Apress, Berkeley, CA, 287-302, (2019).

[46] Winston W., "Microsoft Excel data analysis and business modeling", ***Microsoft press***, (2016).

[47] Gonçalves B. and Sánchez D., "Crowdsourcing dialect characterization through twitter", ***PloS one***, 9.11: e112074, (2014).

[48] Bessho F., Harada T. and Kuniyoshi Y., "Dialog system using real-time crowdsourcing and twitter large-scale corpus", ***In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue***, 227-231, (2012).

[49] Finin T., Murnane W., Karandikar A., Keller N., Martineau J. and Dredze M., "Annotating named entities in Twitter data with crowdsourcing", ***In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk***, Association for Computational Linguistics, 80-88, (2010).

[50] Yadav K., Kumaraguru P., Goyal A., Gupta A. and Naik V.,"SMSAssassin: crowdsourcing driven mobile-based system for SMS spam filtering", ***In Proceedings of the 12th Workshop on Mobile Computing Systems and Applications***,1-6, (2011).

[51] https://drive.google.com/drive/folders/1lQxCokjXov7b WHAMjsXxcrBOunZNX_bW, The Twitter dataset used in this manuscript can be accessed from this link or contact to author, (15 May 2021).

[52] https://developers.virustotal.com/reference#file-search, This is the web page of the virus total site, (13 January 2021)

[53] Gupta A. and Kaushal R.,"Improving spam detection in online social networks", ***In 2015 International conference on cognitive computing and information processing (CCIP)***, IEEE, 1-6, (2015).

[54] Mahmoud T. M. and Mahfouz A. M., "SMS spam filtering technique based on artificial immune system", ***International Journal of Computer Science Issues (IJCSI)***, 9.2: 589, (2012).

[55] Yadav K., Kumaraguru P., Goyal A., Gupta, A. and Naik V., "SMSAssassin: crowdsourcing driven mobile-based system for SMS spam filtering", ***In Proceedings of the 12th Workshop on Mobile Computing Systems and Applications***, 1-6, (2011).

[56] Nuruzzaman M. T., Lee C. and Choi D., "Independent and personal SMS spam filtering", ***11th International Conference on Computer and Information Technology***, IEEE, 429-435, (2011).

[57] Swe M. M. and Myo N. N.,"Fake accounts detection on twitter using blacklist", ***17th International Conference on Computer and Information Science (ICIS)***,IEEE, 562-566, (2018).

[58] https://prospect.io/blog/455-email-spam-trigger-words-avoid-2018/, Some Sensitive words used in social networks, (9 January 2021).

[59] Patil T. R. and Sherekar S. S., "Performance analysis of Naive Bayes and J48 classification algorithm for data classification", ***International journal of computer science and applications***, 6.2: 256-261, (2013).

[60] Genuer R., Poggi J. M. and Malot C., "Variable selection using random forests",***Pattern Recognition Letters***, 31.14: 2225-2236, (2010).

[61] Moradian M. and Baraani A., "KNNBA: K-Nearest Neighbor Based Association Algorithm". ***Journal of Theoretical & Applied Information Technology***, 6.1: (2009).

[62] Boahen E. K., Changda W. And Elvire B. M., "Detection of Compromised Online Social Network Account with an Enhanced Knn", ***Applied Artificial Intelligence***, 34.11: 777-791, (2020).

[63] Kaur G. and Chhabra A., "Improved J48 classification algorithm for the prediction of diabetes", ***International Journal of Computer Applications***, 98.22: (2014).

[64] Rajput A., Aharwal R. P., Dubey M., Saxena S. P. and Raghuvanshi M., "J48 and JRIP rules for e-governance data", ***International Journal of Computer Science and Security (IJCSS)***, 5.2: 201, (2011).

[65] Tapkan P. Z. and Özmen T., "Determining the spam quality by feature selection and classification in a social media", ***Pamukkale University Journal of Engineering Sciences***, 4: 713-719, (2018).

[66] Gerbet T., Kumar A. and Lauradoux C., "A privacy analysis of Google and Yandex safe browsing", ***46th***

ERKEN GÖRÜNÜM

Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), IEEE, 347-358, (2016).

[67] https://www.virustotal.com/intelligence/help, "This is the full list of allowed", (16 May 2021).

[68] Peng P., Yang L., Song L. and Wang G.,"Opening the blackbox of virustotal: Analyzing online phishing scan engines", ***In Proceedings of the Internet Measurement Conference***, 478-485, (2019).

[69] Salem A., Banescu S. and Pretschner A., "Maat: Automatically Analyzing VirusTotal for Accurate Labeling and Effective Malware Detection", ***arXiv preprint arXiv:2007.00510***, (2020).

[70] https://www.virustotal.com/gui/user/oguzhancitlak/apikey, "more about the API functionality in the Virus Total Developer Hub", (16 May 2021).

[71] Witten I. H. and Frank E., "Weka. Machine Learning Algorithms in Java", 265-320, (2000).

[72] Sharma R. C. Hara K. and Hirayama H., "A machine learning and cross-validation approach for the discrimination of vegetation physiognomic types using satellite based multispectral and multitemporal data", ***Scientifica***, (2017).

[73] Dener M., Dörterler M. and Orman A., "Açık kaynak kodlu veri madenciliği programları: WEKA'da örnek uygulama", ***Akademik Bilişim***, 9: 11-13, (2009).

[74] Baskin I. I., Marcou G., Horvath D. and Varnek A., "Cross-Validation and the Variable Selection Bias", ***Tutorials in Chemoinformatics***, 163-173, (2017).

[75] Foozy C. F. M., Ahmad R., Abdollah M. F. and Wen, C. C., "A Comparative Study with RapidMiner and WEKA Tools over some Classification Techniques for SMS Spam", ***In IOP Conference Series: Materials Science and Engineering, IOP Publishing***, 226.1: 012100, (2017).

[76] Xu Q. S. and Liang Y. Z., "Monte Carlo cross validation", ***Chemometrics and Intelligent Laboratory Systems***, 56.1: 1-11, (2001).

[77] Smyth P.,"Clustering Using Monte Carlo Cross-Validation", ***In Kdd***, 1: 26-133, (1996).

[78] https://www.cs.waikato.ac.nz/ml/weka/index.html, "Weka is tried and tested open source machine learning software", (16 May 2021).

[79] Nasukawa T. and Nagano T., "Text analysis and knowledge mining system", ***IBM systems journal***, 40.4: 967-984, (2001).

[80] Baldry A., Thibault P. J., "Multimodal transcription and text analysis", ***London: Equinox***, 26, (2005).

[81] Bozan Y. S., Çoban Ö., Özyer G. T. and Özyer B., "SMS spam filtering based on text classification and expert system", ***23nd Signal Processing and Communications Applications Conference (SIU)***, IEEE, 2345-2348 , (2015).

[82] Colladon A. F. and Gloor P. A., "Measuring the impact of spammers on e-mail and Twitter networks", ***International Journal of Information Management***, 48: 254-262, (2019).

[83] Gloor P. A., Laubacher R., Dynes S. B. and Zhao Y., "Visualization of communication patterns in collaborative innovation networks-analysis of some w3c working groups", ***In Proceedings of the twelfth international conference on Information and knowledge management***, 56-60, (2003).

[84] Bayrakdar S., Yucedag I., Simsek M. and Dogru I. A., "Semantic analysis on social networks: A survey", ***International Journal of Communication Systems***, e4424, (2020).

ERKEN GÖRÜNÜM