Savunma Bilimleri Dergisi, 43(1): 155-177 (2023)

Elektronik/Online ISSN: **2148-1776**

Basılı/Printed ISSN: **1303-6831**

SAVUNMA BİLİMLERİ DERGİSİ

The Journal of Defense Sciences

https://dergipark.org.tr/tr/pub/khosbd

Machine and Deep Learning Studies for Cyberbullying Detection

Siber Zorbalık Tespiti için Makine Öğrenmesi ve Derin Öğrenme Çalışmaları

Mümin Ferhat YAKUT [1] Çağrı ŞAHİN [2] Yılmaz ATAY [3,*]

[1]*Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 06570, Çankaya/ANKARA*

[2] *Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 06570, Çankaya/ANKARA*

[3] *Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 06570, Çankaya/ANKARA*

Makale Bilgisi

Araştırma makalesi
Başvuru: 14.03.2022
Düzeltme: 05.07.2022
Kabul: 27.10.2022

Keywords

Algorithms
Deep Learning
Machine Learning
Cyberbullying
Social Media

Anahtar Kelimeler

Algoritmalar
Derin Öğrenme
Makine Öğrenmesi
Siber Zorbalık
Sosyal Medya

Özet

Toplumdaki internet devrimi, sosyal medya kullanımı gibi günlük hayatımızda çeşitli etkilere sahiptir. Sosyal medya, hayatımızın her alanında kullanılıyor ve bazı alanlarda çok avantajlı olsa da, günümüz dünyasında giderek daha fazla ortaya çıkan yeni bir konuyu da beraberinde getiriyor. Bu yeni konu; Siber Zorbalık, utanç, suçluluk veya aşağılanma duygularına neden olan içerikler göndererek veya paylaşarak birine zarar vermeyi içerir. Sahte kimlikle kolayca sahte sosyal medya hesapları oluşturmak, siber zorbalık olaylarını daha da artırmakta ve siber zorbaları teşvik etmektedir. Siber zorbalık, insanları hem zihinsel hem de fiziksel olarak etkileyebilir ve kalıcı sorunlara yol açabilir. Ancak, bu alanda yapılan çalışmalar siber zorbalığın önlenebilir olduğunu göstermektedir. Bu çalışmada, siber zorbalığı tespit etmek ve önlemek için makine öğrenmesi tekniklerini gözden geçiriyor, makine ve derin öğrenme modellerinin performanslarını değerlendiriyor ve modellerin performansını etkileyen faktörleri inceliyoruz. Ayrıca, siber zorbalık tespitinde veri ön işleme, sınıflandırma, öznitelik çıkarma ve seçme süreçlerinin önemini tartışıyoruz. Böylece siber zorbalığın tespiti ve önlenmesi konusunda çalışacak araştırmacılara öznitelik çıkarma yaklaşımları, özellik seçme teknikleri ve sınıflandırıcıların seçimi konularında genel perspektif kazandırılmıştır. Ayrıca metinsel verilerin yanı sıra fotoğraf, video ve ses verileri üzerinde siber zorbalık tespit çalışmalarının yapılabileceği de vurgulanmıştır.

Abstract

The internet revolution in society has various effects on our daily life such as the use of social media. While social media is ubiquitous and great in some aspects, it brings a new issue that appears more and more in today's world. This new issue, cyberbullying, involves harming someone by posting or sharing content that causes feelings of embarrassment, guilt, or humiliation. Easily creating fake social media accounts with fake identity further increases cyberbullying incidents and encourages cyberbullies. Cyberbullying can affect people both mentally and physically and can lead to permanent problems. However, studies in this area show that cyberbullying can be prevented. In this study, we review machine learning techniques to detect and prevent cyberbullying, evaluate the performances of the machine and deep learning models, and examine factors that affect the performance of the models. We also discuss the importance of data preprocessing, feature extraction and selection, and classification processes in cyberbullying detection problems. Thus, a general perspective on feature extraction approaches, feature selection techniques, and selection of classifiers have been given to researchers who will work on the detection and prevention of cyberbullying. It is also emphasized that cyberbullying detection studies could be carried out on photographic, visual and audial data as well as textual data.

* Yılmaz ATAY, e-mail: yilmazatay@gazi.edu.tr
DOI: 10.17134/khosbd.1087548

1. INTRODUCTION

The internet has become part of our daily life with various usage areas and web-based applications. The most popular way to use Internet is social media that people from all ages involved to share some contents. Based on the recent studies, social networks such as Facebook have billions of users in worldwide (Bozyigit, Utku and Nasibov, 2021; Manca, Bocconi and Gleason, 2021). While social media provides users having undeniable benefits such as direct, fast, easy communication with people in the social network, it might be harmful at the same time (Yin et al., 2009). Cyberbulling, harming someone by posting or sharing contents that cause feelings of embarrassment, guilt, or humiliation via digital technologies, is an example of how social media can be threat to its users (Bauman, Toomey and Walker, 2013; Nandhini and Sheeba, 2015; Hemphill, Kotevski and Heerde, 2015). Since people can obtain fake social media accounts with fake identity, it is easy to reach lots of people in a short time and insufficient legal regulations, cyberbullying on social media increases rapidly. Victims of cyberbullying can be a person, group of people, or organization (Rosa et al., 2019). Victims who are exposed to Cyberbullying can have more lasting and profound effects compare to traditional bullying especially for children (Dadvar and De Jong, 2012). Low self-esteem, depression, and even suicide are some of the emerging consequences of Cyberbullying (Smith et al., 2008). Therefore, detection and prevention of this threat have become a major concern for researchers. To detect and prevent Cyberbullying, researchers rely on different approaches such as supervised-based learning, lexicon-based learning, rules-based learning, and hybrid-based approaches (Salawu, He and Lumsden, 2017). While supervised learning-based approaches leverage traditional machine learning algorithms to develop predictive models, lexicon-based approaches use word lists and the presence of words in the lists for detection (Muneer and Fati, 2020; Reynolds, Kontostathis and Edwards, 2011). In the rules-based approaches, text matching with predefined rules is considered to identify bulling (Perera and Fernando, 2021). Hybrid approaches use combination of one or more of the existing approaches with human-based reasoning (Bozyigit et al, 2021).

In this paper, we review the effectiveness of machine and deep learning algorithms in detecting and preventing cyberbullying. The remainder of the paper is organized as follows: Section 2 defines the problem of cyberbullying. Sections 3 and 4 present the literaure review and the methods used in the existing studies, respectively. Finally, Section 5 presents our conclusions and future work.

1.1 Problem

Excessive use of social networks at all ages causes an increase in cyberbullying incidents. Especially after the Covid-19 pandemic, children and adolescents have most affected by cyberbullying due to the increase in the time they spend on the internet. Recent studies show that 36.5% of people are exposed to cyberbullying (Cheng, Silva, Hall and Liu, 2021). Moreover, according to a 2019 study by Rao et al., 44.5% of young people in China have been victims of cyberbullying. (Rao et al, 2019) According to statistics released by the Office of National Statistics, 7 out of 10 children are emotionally affected by

cyberbullying. 20% of children have been directly exposed to cyberbullying. Detecting and preventing cyberbullying on social media has become increasingly important in recent years, as victims of cyberbullying can experience very strong reactions that can include low self-esteem, depression, and even suicide.

1.1.1 Definitions

While traditional bullying is defined as "intentional harmful behavior that involves an imbalance of power and results in repeated, aggressive behavior" (Chudal et al., 2021), cyberbullying is defined as a type of traditional bullying that is generally done using information technologies. For example, Kowalski and Limber (Kowalski and Limber, 2007) defined cyberbullying as bullying with messages sent via a website or mobile phone, while Hinduja and Patchin (Hinduja and Patchin, 2008) emphasized that cyberbullying is carried out continuously, deliberately and through electronic devices with the aim of causing harm. However, a comprehensive definition of cyberbullying is very important as it will form the basis of the studies to be done in this field and there is a need for a worldwide standard definition. Chun et al. (Chun, Lee, Kim and Lee, 2020) examined different definitions of cyberbullying in their study and stated the most commonly used features when defining cyberbullying as follows. Use of electronic means (100%), repeated harm or behavior (32.8%), deliberate or intentional act (31.3%), vulnerability (15.6%), unwanted information of others (3.1%) and purpose for threatening, harassing, or embarrassing others.

1.1.2 Types of cyberbullying

There are many different types of cyberbullying compared to traditional bullying (Nadali, Murad, Sharef, Mustapha and Shojaee, 2013). Commonly known types of cyberbullying are ostracism, harassment, disclosure of confidential information, cyberstalking, deception, trolling or outright insults and swearing. With the widespread use of social networks, bullies can easily and effectively victimize people on issues such as sexuality, harassment, threats, exclusion, appearance and racism (Perera and Fernando, 2021; Talpur and O'Sullivan, 2020). Considering the effects of cyberbullying cases on victims, it can cause significant and permanent problems emotionally, mentally and physically**.**

1.1.3 Motivation

Cyberbullying has become a threat especially for children and young people. This threat is growing day by day parallel with the increase in number of social media users. With the increasing role of technology in our daily lives, the number of victims of cyberbullying has also increased. In addition to this, studies on detecting and preventing cyberbullying have also gained importance. According to Google Trends data, research on cyberbullying around the world has increased 4 times from 2004 to present. The real-life consequences of cyberbullying incidents result in cases such as suicide, lack of self-confidence, depression, and mental health disorders according to Miller's study (Miller, 2017). Consequently,

cyberbullying events, the effects of which are understood more clearly over time, have been found remarkable by researchers and studies have been carried out to detect and prevent cyberbullying. This study is presented to provide information for researchers who want to work in the field of cyberbullying. The studies examined are studies that include ML and DL approaches and have been examined according to years.

2. LITERATURE REVIEW

Initial studies on cyberbullying detection are generally aimed at the effectiveness of the textual, contextual, user-based, and social networks features. Yin et al. (Yin et al., 2009) presented one of the earliest study on the automatic detection of cyberbullying by using machine learning algorithms. They collected three different datasets from chat-style communities (Kongregate) and discussion-style communities (Slashdot and Myspace), and used Term Frequency-Inverse Document Frequency (TF-IDF) and N-Gram techniques for feature extraction. According to the experimental study results, the use of TF-IDF and contextual features increased the success of the classifier. However, the results obtained are not sufficient. Because, according to the test results made by combining both feature extraction, the F-Score was measured as 0.44. However, the study played an important role for future studies in terms of creating a dataset on online harassment detection and drawing attention to this problem.

Reynolds, Kontostathis, Edwards (Reynolds et al., 2011) examined the question-answer dataset obtained from Formspring.me. In this study, C4.5, JRIP, Instance Based (IBK) and Sequential Minimal Optimization (SMO) algorithms are used through the Weka application. The developed model focused on text features. A bad words list containing 296 bad words was used and these words were leveled. While labeling the data, Amazon's Mechanical Turk service was used. Recall has been determined as a performance measurement tool. SMO gave the worst performance. The most successful results were given by the IBK and C4.5 algorithms.

Dadvar et al. (Dadvar, De Jong, Ordelman and Trieschnigg, 2012) investigated the effect of gender-specific language features on the detection of cyberbullying using the dataset collected from the Myspace (https://myspace.com). Their results show that the use of gender-specific language features provide better success than basic methods for the classification. In addition to this study, Dadvar et al. (Dadvar, Trieschnigg, Ordelman and De Jong, 2013) worked on the dataset containing comments and user information collected from Youtube. The classifier is trained with content-based, cyberbullying, and user-based features. As a result, the use of cyberbullying and user-based features along with Content-based features has increased the success of classification.

Qianjia et al. (Huang, Singh and Atrey, 2014) proposed a composite model using textual features and social network features and they observed that the use of social network features enhancing the success of the classifier. Al-Garadi et al. (Al-garadi, Varathan and Ravana, 2016) proposed a machine learning model for the detection of cyberbullying in 2016 with a unique set of features obtained from Twitter. These features include network, activity, user and tweet content. Three different algorithms Chi-Square,

Information Gain and Pearson Correlation were used for feature selection. A balance was achieved between the classes in the dataset with SMOTE and cost-sensitive techniques. They applied NB, SVM, RF and KNN algorithms for classification with different parameters and compared the results with AUC (Area Under the ROC Curve) and F-Score values. When the proposed method was applied with SMOTE and Random Forest algorithms, it reached the most successful result with 94.3% AUC and 93.6% F-Score values. This showed that SMOTE technique was more efficient than cost-sensitive technique for this dataset.

Kargutkar and Chitre (Kargutkar and Chitre, 2020) used the CNN algorithm to detect cyberbullying. Since the CNN algorithm uses numerical inputs, they generated word vectors after cleaning the dataset. Their designed model consisted of three layers: data preprocessing, CNN model layer and model prediction layer. The ReLu function was used as the activation function in the CNN model. The results show that the designed model is more successful than traditional machine learning models. Ozel, Sarac, Akdemir and Aksu (Ozel, Sarac, Akdemir and Aksu, 2017) applied SVM, C4.5, MNB and KNN algorithms to Turkish texts they collected on Instagram and Twitter with Information Gain and Chi-Square feature selection methods. Emojis in the dataset were removed from the data in the preprocessing stage, and a second dataset was obtained and studies were carried out on these two data. An emoji list was prepared manually for the emojis in the dataset. The stemming process was not applied to the dataset as it was not considered to be suitable for the Turkish language. In addition, 5-fold cross-validation was applied while the dataset was separated as training and test dataset. In the experimental results obtained, the inclusion of emojis in the dataset and the application of feature selection techniques increased the success of the classifier. The Information Gain technique provided more successful results than the Chi-Square technique. When the running times of the machine learning algorithms applied on the dataset were also included in the evaluation, the MNB algorithm showed the most successful performance.

Sahni and Raja (Sahni and Raja, 2018) applied the sentiment analysis method with machine learning classifiers for the detection of cyberbullying on Hindi and English texts and evaluated the results. Twitter data was used as the dataset. In the experimental study, NB, RF and J48 classifiers were tested. In addition, classification was made using sentiment analysis techniques and the results were compared. According to the results, all three of the machine learning algorithms showed a high performance of 98%. Accuracy and Kappa statistics were used for the performance evaluations of the algorithms.

Altay and Alatas (Altay and Alatas, 2018) performed performance measurements using natural language processing techniques and machine learning models such as Bayesian Logistic Regression (BLR), RF algorithm, Multilayer Perceptron, J48 and SVM algorithms. The dataset was obtained by Sergio Jiménez Barrio from the data on the Formspring.me site. In the data preprocessing stages, stemming, tokenization and removal of unnecessary words were applied. In feature extraction, TF-IDF and document term matrix were used. In the experimental study, the results obtained by separating the test and training data on the dataset or by applying 10-fold cross-validation techniques were compared. According to the

results obtained, the RF algorithm gave the most successful result by using the entire dataset as training data.

Hussain, Mahmud and Akthar (Hussain, Mahmud and Akthar, 2018) also collected dataset from Facebook, Prothom-Alo news and YouTube comments for the detection and prevention of cyberbullying. While labeling the dataset, each comment was evaluated by at least fifty people and class labeling was performed according to these results. Special characters and punctuation marks are removed in data preprocessing stages. N-Gram was used as the feature extraction method. In the study, a method is proposed in which the weight values are calculated according to the occurrence of words in documents. In other words, for each word, its weight is calculated according to the documents labeled as cyberbullying and according to the documents labeled as cyberbullying-free. The weight values of the words were used while making the classification.

Al-Mamun and Akhter (Abdullah-Al-Mamun and Akther, 2018) tested SVM, NB, J48 and KNN machine learning models on the Bangala language dataset they collected from Facebook and Twitter. As a result of performance measurements, the SVM model showed the most successful result with a very high accuracy rate of 97%. According to the results obtained, it has been shown that the inclusion of eleven basic information such as location, age, gender about the users together with the posts of the users positively affects the result. In addition, in this study, a performance comparison of the classifications made on English texts and the classifications on Bangala texts was made. SVM was the most successful classifier in both English and Bangala.

Sintiha and Mostakim (Sintaha and Mostakim, 2018) aimed to compare different approaches for detecting cyberbullying in their study. For this, they applied NB and SVM models to the dataset they collected from the Twitter environment on the detection and prevention of cyberbullying on social media. The dataset was collected using some keywords via the Twitter API. In addition, since the emojis that are frequently used in the posts are very effective in conveying emotions, they were included in the analysis by matching with certain codes instead of being removed from the posts. According to the performance measurements, SVM gave more successful results than the NB algorithm with accuracy 89,54%.

Bozyigit et al. (Bozyigit, Utku and Nasiboglu, 2018) emphasized that among the detection studies conducted in the field of cyberbullying, those focused on Turkish texts are very few and there is no Turkish dataset in the field. For this reason, a dataset from social media comments with Turkish content was prepared and presented for the use of researchers. In the data preprocessing stage on this dataset, a dictionary containing 144 Turkish insults and swearing and Levenshtein distance were used to correct the words with spelling errors. Thus, incorrect words in the dataset were removed. Bag of Words technique was used together with TF-IDF and Information Gain technique for feature extraction. NB, SVM, RF, KNN, MNB and C4.5 algorithms were tested and performance measurement was made with

the F-Score value. Considering the runtimes and F-Score values of the classifiers, the most successful methods were found to be MNB, SVM and KNN.

Al-Ajlan and Ykhlef (Al-Ajlan and Ykhlef, 2018a) developed a deep learning method called CNN-CB which ignores the feature extraction and feature selection steps used in traditional cyberbullying detection methods. The large dataset causes the number of features to increase and feature selection becomes difficult. Classification based on the features determined as a result of feature selection ensures that the structure is static. Architectural structure of the dynamic CNN-CB algorithm consists of four layers: word embedding, convolutional, max pooling and dense layer. In the first layer, word embeddings are created for each word in the tweet. Then the vectors obtained from this layer are sent as input to the second layer. In the second layer, the input vectors are compressed without losing important features. The third layer takes the output of the second layer as input and finds the maximum value of the selected region to record only important events. In the last layer, classification is done. In the experimental study, the dataset was obtained using the Twitter API. Obtained data were first cleared of duplicate records containing only images and URLs and labeled. The CNN-CB algorithm provided results an accuracy value of 95%.

Al-Ajlan et al. (Al-Ajlan and Ykhlef, 2018b) proposed the use of word embedding as a new method. In the first step of their experimental study, 20,000 tweets were randomly collected using the twitter4j API and java code. They removed repetitive and irrelevant data from collected tweets and relied on The GloVe model to capture the similarities between words. In addition to these, metaheruistic optimization algorithms were used to optimize the parameters of the CNN algorithm in the classification phase. They concluded that updating the parameters of the CNN algorithm to optimal or close to optimum values greatly improves the classification results.

Curuk, Acı and Essiz (Curuk, Acı and Essiz, 2018) aimed to detect and prevent cyberbullying online. In their studies, they used the comments obtained from Formsprig.me and Myspace environments as a dataset. They aimed to improve the success and performance criteria by removing stopwords in the data preprocessing stage. For feature extraction, N-Gram technique was used and n=1. In addition, the TF-IDF technique was used to measure the weights of the words. In the experimental study, artificial neural network based SVM, SGD, Radial Based Function (RBF) and LR classifiers were tested. The F-Score value was used for performance measurement and the SGD method provided the highest success with 95% for the Formspring.me dataset. For the Myspace dataset, each of the classifiers showed a high success rate of 98%. Haidar et al. (Haidar, Chamoun and Serhrouchni, 2018) presented a feed forward neural network (FFNN) model for cyberbullying detection. They updated and used The Arabic dataset presented an existing work (Haidar, Chamoun and Serhrouchni, 2017). The dataset is divided into "small dataset" and "large dataset". In the experimental study, a feedforward neural network with four hidden layers was used. According to the results obtained, the feedforward neural network gives more successful results in large datasets.

In their study, H. Rosa et al. (Rosa et al., 2019) presented a comprehensive experimental study with literature review for the detection and definition of cyberbullying. In the experimental study, 12 different scenarios were tested by using two different datasets, Bullying Traces Dataset and Formspring, with different feature extraction scenarios. The most successful classifier for the first dataset was the SVM model, which was implemented with scenario A: TF-IDF, scenario D: TF-IDF + Word Embeddings, scenario H: TF-IDF + Personality Trait Features + Word Embeddings. Here, F-Score was used as a performance measurement criterion and a success rate of 0.74 was obtained. For the second dataset, the most successful classifier was again the SVM implemented with "scenario H". In this application, F-Score=0.45.

In their study, Balakrishnan, Arabnia and Khan (Balakrishnan, Khan and Arabnia, 2020) obtained 5453 shares on Twitter by using the #Gamergate hashtag and used them as a dataset. NB, RF and J48 classification algorithms were used for classification. The collected dataset was first divided into four classes (bully, aggressive, spam and normal) by experts and labeled. In addition, user personalities were determined by using Big Five (Openness, Conscientiousness, Extraversion, Adaptability and Neuroticism) and Dark Triad (Machiavellianism, Narcissism, Psychopath) models and a relationship was tried to be established between users' personalities and cyberbullying detection. In addition to the personalities of the users, the effect of their emotions on cyberbullying was tried to be measured. Indico API was used in emotion analysis and the categories of anger, fear, joy, sadness and surprise were examined. However, it has been observed that the cyberbullying situation is more affected by the personalities of the users than their emotions. A 10-fold cross-validation technique was used to prevent the trained classifiers from predicting some classes better and some worse due to the disproportionate distribution according to the classes in the dataset and to increase the robustness of the classifiers. According to the results obtained, it was seen that user personalities improved the detection of cyberbullying. In addition, RF and J48 algorithms have obtained more successful results than NB.

Bozyigit, Utku and Nasiboglu (Bozyigit, Utku and Nasiboglu, 2019) used Artificial Neural Networks (ANN) model to detect cyberbullying in Turkish Twitter posts. Using TF-IDF and N-Gram for feature extraction, the researchers also edited the abbreviations and spelling mistakes in the shares during the data preprocessing stage. For this arrangement, firstly, more than two repetitive letters in a word were removed, and then a similarity study was carried out based on a list of 144 words, which are considered as insults and curses in Turkish. Thus, incorrect words were corrected. Eight different ANN models were designed for learning and F-Score value was used for performance measurement. The highest success was obtained in the 2-layer neural network with 91%.

In 2019, John Hani et al. (Monir et al., 2019) proposed a machine learning approach for cyberbullying detection and prevention. In the data preprocessing step, there are tokenization, converting texts to lowercase, removing unnecessary words and finding the correct spellings of words. TF-IDF technique was used for feature extraction. Sentiment analysis technique was also used to classify the post as

positive or negative and to add the obtained words to the TF-IDF word list. In addition, the N-Gram technique was used to evaluate different word combinations. SVM and Neural Network (NN) models from machine learning algorithms were used as classifiers. In neural networks, there are 128 nodes in the input layer and 64 nodes in the hidden layer. According to the results obtained, the NN model showed the highest performance with 92.8% in the 3-gram technique, while SVM showed the highest performance in the 4-gram technique with 90.3%.

Banerjee et al. proposed a new model for detecting cyberbullying using word vectors and CNN algorithm (Banerjee, Telavane, Gaikwad and Vartak, 2019). To extract word vectors, they applied GloVe technique. The word vectors obtained from the first layer were used as inputs for the CNN in the second layer. In the third layer, the maxpooling layer, the most valuable features were selected. The proposed model was tested on Twitter data and it was seen that 93.97% success was achieved.

Sudhanshu Dalvi, Baliram Chavan and Halbe (Dalvi, Chavan and Halbe, 2020) used the data obtained from Twitter via the Tweet API in their study in 2020. After the collected data is passed through the data preprocessing steps, feature extraction is performed with the TF-IDF technique, and after this step, SVM and NB classifiers are used to calculate the probabilities of the tweets received to decide whether a tweet contains bullying. If the calculated probability value is less than 0.5, it is decided that there is no bullying. However, if the probability value is greater than 0.5, the last ten tweets of the same user are fetched again and the probability is calculated again. If the average probability values obtained are greater than 0.5, this post is labeled as bullying and it is determined that this user is bullying. Natural Language Toolkit (NLTK) was used for tokenization in the data preprocessing stage. In addition, stopwords that do not affect the result have been removed and all texts have been converted to lowercase. Synonyms were identified and singularized. TF-IDF technique was used for feature extraction. The main reason for using SVM and NB as classifiers is that these classifier models calculate probability for each class. According to the results obtained, the SVM model was the most successful classifier.

Muneer and Fati (Muneer and Fati, 2020) compared the performances of seven different machine learning models for the detection of cyberbullying on the data they collected from the Twitter environment. In the study, punctuation marks, stopwords (unnecessary words, conjunctions, etc.) and special characters were removed in the data preprocessing stage, and stemming was done for words. Then, TF-IDF and/or Word2Vec were used for feature extraction. The systems trained using the obtained features and machine learning models were tested in the test data. Considering the accuracy, precision, recall and F-Score values of the classifiers used, the most successful algorithm is LR. The Stochastic Gradient Descent (SGD) and Light Gradient Boosting Machine (LGBM) classifiers were also able to extract accuracy and F-Score values close to LR. As a result of the experimental study, it was seen that the LR method gave better results as the data size increased.

Bandeh Ali Talpur and O'Sullivan (Talpur and O'Sullivan, 2020) on the dataset consisting of approximately 50,000 tweets obtained from the Twitter environment, after the preprocessing steps,

feature extraction was performed with different methods (POS tagging, PMI, Emotional, Dictionary, etc.). SVM, K-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), Naive Bayes (NB) algorithms were used and the performance values of these algorithms were compared. In the study, it was aimed to determine the type and severity of cyberbullying. Therefore, a dictionary containing words related to types of cyberbullying (sexuality, racism, physical appearance, intelligence and politics) was used. In order to classify tweets according to their severity, 4 classes were created as high, medium, low and non-bullying. Here; High violence=sexuality and physical appearance, Moderate violence=politics and racism, Low violence=intelligence. It has been seen that the RF algorithm is more successful than the others. Since the balanced distribution of the labeled data according to the classes will affect the classification success, the SMOTE technique was used and it was observed that the success rate increased. In addition, the richness of the words related to the topics in the dictionary used also affects the success rate of the classifier.

Ximena, Cuzcano and Ayma (Cuzcano and Ayma, 2020) compared the performances of NB, Multinomial Logistic Regression, SVM and RF methods using Spanish Twitter data. In the study, the classification was made according to four different categories as "non-harassment", "direct harassment", "hate speech" and "sexual harassment". 10,096 tweets collected from Twitter with the help of the Streaming API were labeled according to these four categories by undergraduate students from different universities in Peru. In the study, a tweet was labeled by at least three different students to determine the class label. TF-IDF, N-Gram were used for feature extraction. The most successful model was determined as SVM.

On and Yeniterzi (On and Teniterzi, 2020) tested Convolutional Neural Networks (CNN), SVM, MNB and BLR models on the detection and prevention of cyberbullying in Turkish Twitter posts. Word vector representations were used to express the relations of words, spelling and semantic features. Information sent with pre-trained word vectors achieves better performances. Word2Vec and FastText methods were used to create these vector representations. In the study, besides the CNN models created with three different word vector representations, the CNN model created without using any word vector representation was also considered. A single layer CNN model was designed and the Sigmoid function was used as the activation function. F-Score was used for performance evaluation. As a result of the experimental study, it was seen that more successful results were obtained when the random word vector representation was used. This situation reveals the importance of word vector representation study.

The study of "Cyberbullying detection solutions based on deep learning architectures", it has been evaluated the performance of deep learning methods for detecting cyberbullying. In the data preprocessing stage, stemming, text cleaning, tokenization and Lemmatization were applied. They applied Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) algorithms on the dataset and they obtained the most successful results with BLSTM (Iwendi, Srivastava, Khan and Maddikunta, 2020).

Perera and Fernando mentioned that cyberbullying is deliberate and causes permanent damage on people (Perera and Fernando, 2021). Therefore, they evaluated TF-IDF, sentiment analysis, and profanity filter for feature extraction on Twitter dataset. Support Vector Machine (SVM) was used for classification and Logistic Regression (LR) was used to find the best feature combination. Based on their results, TF-IDF and sentiment analysis combination gives better results. In another study, Alsubait and Alfageh (Alsubait and Alfageh, 2021) compared the performances of Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB) and LR methods on comments collected from famous Arabic channels on Youtube. 15.050 shares were collected from famous Arabic youtube channels and the posts were labeled by three researchers according to whether they contain cyberbullying or not. During the data preprocessing stages, punctuation marks, symbols, URLs, hashtags were removed. All documents have been converted to lowercase letters and stemming has been done for the words. In addition, two different feature extraction techniques, Count Vectorizer and TF-IDF Vectorizer were used. According to performance measurements, it has been seen that LR methods are more successful when Count Vectorizer is used, and CNB methods are more successful when TF-IDF Vectorizer is used. Manowarul Islam et al. (Islam et al., 2021) compared the performance of various machine learning algorithms for cyberbullying detection. They relied on BagOfWords and TF-IDF techniques as feature extraction and collected datasets from Facebook and Twitter. In the experimental study, the SVM algorithm showed more successful results when used with both BagOfWords and TF-IDF techniques in both datasets. In a study conducted in 2021, Bozyigit et al. presented an observation that the features of users and posts play an important role to detect cyberbullying in addition to the textual data collected from Twitter. In the data collection phase of the study, the lack of a detailed Turkish dataset for cyberbullying studies was emphasized and information about tweets and users was collected via Twitter API. The collected data are shared in (Bozyigit, 2020) for further studies in this area. In order to create a balanced dataset, the classifier proposed in (Bozyigit et al., 2019) was used in the data elimination step. The crowdsourcing approach was used in the labeling of the data. The Chi-Square technique was used to measure the effect of social media features on the classification process. According to the Chi-Square test results, Retweets, Favorites, SenderFollowers and SenderLocation features played a more important role in determining the classes. In addition, it is aimed to prevent machine learning algorithms from being affected by the abnormal distribution of these values by applying min-max normalization for social media features with numerical values. Spelling mistakes in Tweets have been tried to be corrected using the slang words list used in (Bozyigit et al., 2019) and Levenshtein Distance. Since it was considered not suitable for Turkish words, stemming was not applied. The application of min-max normalization to social media features has considerably reduced the running times of machine learning algorithms. BagOfWords and TF-IDF approaches were used in feature extraction steps. The dataset was prepared in two different variations. While only textual features were used in the first variation, both textual and social media features were used in the second variation. Grid search was applied to determine the optimal parameters to be used in machine learning algorithms. Since it is considered that a balanced

dataset has been created, accuracy was primarily taken into account in performance measurement. As a result of the experimental study, it was seen that social media features increased the classification success. The most successful result was the Adaptive Boosting (AdaBoost) algorithm in both dataset variations.

Bharti et al. tested machine learning and deep learning algorithms on the dataset obtained from Twitter (Bharti, Yadav, Kumar and Yadav, 2021). In the study, BagOfWords for machine learning algorithms and word embedding obtained with different versions of GloVe technique for deep learning algorithms were applied. They made Perfromas evaluation via Accuracy, Precision, F-Score, Recall and AUC metrics. Based on the results, the most successful machine learning model was LR with an F-Score value of 94.19%, while the most successful deep learning model was BLSTM applied with GloVe 840 with 94.20%. Choi, Jeon and Kim (Choi, Jeon and Kim, 2021) examined about 65,000 posts in their study on the Daum Agora community used in Korea. In this study, text mining and social network analysis were used together. In the study, while deciding whether the posts are harmful or not, attention was paid to whether they contain derogatory words or not. For this, a dictionary of derogatory words was created and used in this study. Not only the cyberbullying user was detected, but also how active the detected user was in this community was evaluated with social network analysis methods. For this evaluation, first of all, users' shares, comments, etc. A graph was designed with this in mind. With the help of this graph, the cyberbullying score of the users was calculated.

Murshed et al. (Murshed et al., 2022) proposed a new method called DEA-RNN for cyberbullying detection by combining the DEA (Dolphin Echolocation Algorithm) optimization algorithm and the Elman type RNN approach. In the developed method, RNN parameters are automatically arranged with DEA. In the study, various pre-processing steps (punctuation, hashtags, symbols and stopwords removed from dataset and all records are converted to lowercase) were applied on the data set obtained from the Twitter platform. In addition, Word2Vec, TF-IDF, POS tagging and Information Gain techniques were used in the feature extraction stage. Spelling mistakes in the data set have been arranged as much as possible. Due to the unbalanced class distribution in the data set, the SMOTE technique was also used. The proposed DEA-RNN method has been compared with different deep learning (Bi-LSTM, RNN) and machine learning (SVM, MNB and RF) models. Accuracy, precision, recall, F-Score and Specificity values were used to evaluate the results obtained. The classifiers were run 20 times and the results were compared by averaging the evaluation metrics. The proposed DEA-RNN method was the most successful with an average accuracy of 90.45%.

3. MATERIAL AND METHODS

In this section, we give information about datasets, preprocessing steps, algorithms, feature extraction and feature selection methods used in reviewed studies that aim to detect and prevent cyberbullying. It has been presented a summary of the studies that leverages machine learning methods in Table 2. The first column, “Study”, lists studies in the literature and their year of publication. The second column,

Dataset, indicates dataset's platform and language for each study. Here, the platform simply represents the social network from which the dataset was collected. The techniques used in the data collection steps and the content of the collected data differ according to the studies. This is one of the reasons why the success rates of the studies are different. The remaining columns provide a brief information about feature techniques, evaluation metrics, classifiers used in the studies and which classifier/classifiers are best among others. The only difference between Table 2 and 3 is that Table 2 lists a summary of the studies that leverages deep learning methods. Tables 2 and 3 are intended to provide a general summary of each article reviewed in this study. Table 2 and 3 show that deep learning algorithms have been used frequently in cyberbullying detection studies in recent years and very successful results have been obtained.

3.1 Data Collection and Labeling

To detect and prevent cyberbullying, collecting and labeling datasets is the primary requirement. The most common tools used to collect required data are Instagram API, Twitter Streaming API, twitter4j API or specially developed software applications are the most common tools used to collect data [Cuzcano and Ayma,2020; Sintaha and Mostakim, 2018; Bozyigit et al., 2018]. Researchers have generally created their own datasets using such tools from social media environments. Therefore, more successful results can be obtained with different classifiers. Data labeling is the labeling process of the records in the dataset, can be used in supervised learning (Muneer and Fati, 2020). Since labeling is usually done manually by experts or users, it is a costly process [Nandhini and Sheeba, 2015; Hemphill et al., 2015]. In particular, some deep learning algorithms generate features automatically, although this may benefit the feature extraction stage, which may require more labeled data (Roh, Heo and Whang, 2021).

3.2 Preprocessing

Preprocessing steps are necessary to create reliable and convenient datasets and reduce the costs such as as running time (Bozyigit et al., 2019). The most common preprocessing steps applied on datasets in the considered studies are; punctuation marks, hashtags, symbols, stopwords, URLs and special characters are removed from the dataset and all records are converted to lowercase (Muneer and Fati, 2020; Curuk et al., 2018; Dalvi et al., 2020). Stemming should be preferred considering the language features used (Bozyigit et al., 2021; Muneer and Fati, 2020; Alsubait and Alfageh, 2021).

Ozel and Sarac (Ozel and Sarac, 2017) evaluated the effect of some feature extraction and selection techniques in detecting cyberbullying. Different combinations of tokenization, stop words removal, stemming and lowercase techniques were tested on the Formspring.me dataset. In addition, Information Gain and Chi Square techniques are tested for feature selection. In order to see how the changes made in the preprocessing phase affect the success of the classifier on the train dataset, four different classifiers were studied with the test dataset. According to the results obtained, stemming and stop words removal

reduced the success of the classifier for the dataset used. However, it is clear that the preprocessing steps to be used for a study may vary depending on the dataset and the language in which this dataset is prepared.

3.3 Feature Extraction and Selection

Learning to classify text in large volumes of data is quite difficult since there are huge numbers of different words and terms (Alam, Bhowmik and Prosun, 2021). In the considered studies, we observed that the inclusion of all features in the study increased the running time of the designed model (Bozyigit et al., 2021). More successful results can be obtained in less time by using fewer features as reported (Dadvar and De Jong, 2012; Balakrishnan et al., 2020). Therefore, determining the most efficient features in the classification process by using various techniques in the Feature Extraction and Feature Selection steps is one of the major focus.

Feature Extraction techniques are classified under three categories which are User-Based Features, Text-Based Features and Activity-Based Features (Al-garadi et al., 2016). While User-Based Feature attributes are obtained from common user profiles such as age, gender, session statistics, Activity-Based Feature attributes is obtained from social networking applications like popularity, Influence, Reciprocity, Power Difference, Centrality Scores, Hubs and Authority, Communities. Text-Based Features refer to the attributes in the text content base on the sentiment, opinion, hate and profanity words and so on. Most of the features obtained after the Feature Extraction steps are useful for classification. However, some features may not be useful and can be excluded from the modeling analysis (Chatzakouy et al., 2017). Feature Selection techniques can be categorized as Filter-Based Feature Ranking Techniques, Filter-Based Feature Subset Techniques, and Wrapper-Based Feature Subset Techniques (Ghotra, McIntosh and Hassan, 2017).

3.4 Algorithms

The machine learning algorithms are mostly used in the considered studies to detect cyberbullying. In addition to that, deep learning algorithms provide succesful results in some studies (Al-Ajlan and Ykhlef, 2018a; Banerjee et al., 2019). When the studies are examined, the most widely used algorithms are the algorithms of SVM, NB, RF and LR as machine learning algorithms, and the approaches of CNN, BLSTM as deep learning methods for detecting cyberbullying (Ozel et al., 2017; Bozyigit et al., 2018; On and Yeniterzi, 2020).

3.5 Evaluation Metrics

In the experimental studies on cyberbullying detection and prevention, classifier models are designed and the performance of the designed models are evaluated. The most common metrics used in performance evaluation are accuracy, precision, recall, and F-Score (Muneer and Fati, 2020; Dalvi et

al., 2020; Monir et al., 2019). Also, Kappa Statistics and AUC are used in some studies (Sahni and Raja, 2018; Al-garadi et al., 2016).

The most common source used to measure classifier success is the confusion matrix. Accuracy, precision, recall and F-Score are derived from confusion matrix. In the studies presented in Tables 2 and 3, the F-Score value was generally used to compare the classifier success. However, in some cases, these metrics may be insufficient. The AUC value, which expresses the area under the ROC (Receiver Operating Characteristic) curve, is more commonly used, especially when working with data sets where the class distribution is unbalanced. ROC is a probability curve for different classes. Kappa Statistic is widely used in both unbalanced and multi-class datasets. Table-1 provides descriptions of these evaluation metrics.

Table 1: Description of Evaluation Metrics

Metrics	Equation	Description
Accuracy	$\frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}}$	A measure of how often the classifier method makes accurate predictions.
Precision	$\frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}}$	It is the division of correctly estimated values by the total values.
Recall	$\frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$	It shows how successful the positives.
F-Score	$\frac{2 * \boldsymbol{Precision} * \boldsymbol{Recall}}{\mathbf{Precision} + \mathbf{Recall}}$	It is the harmonic mean of the Precision and Recall metrics.
Kappa	$\frac{\boldsymbol{P_{observed}} - \boldsymbol{P_{bychange}}}{\mathbf{1} - \boldsymbol{P_{bychange}}}$	It is obtained using the $\boldsymbol{P_{observed}}$ (Probability of observed agreement) and $\boldsymbol{P_{bychange}}$ (Probability of agreement by change).
AUC	Area under the ROC Curve using TPR (True Positive Rate) and FPR (False Positive Rate).	

Table 2. The Summary of Machine Learning Approaches based on Cyberbullying

Study	Dataset		Method And Techniques			
	Platform	***Language***	***Feature Techniques***	***Evaluation Metrics***	***Classifiers***	***Best Classifier***
(Bozyigit et al., 2021)	Twitter	Turkish	BagOfWords, TF-IDF	Accuracy: %90.1 Precision: %90.4 Recall: %88.4, F-Score: %89.4	SVM, LR, KNN, MNB, AdaBoost, RF	AdaBoost
(Islam et al., 2021)	Facebook, Twitter	English	BagOfWords, TF-IDF	Accuracy: %76 (Dataset-1) Accuracy: %80 (Dataset-2)	DT, NB, SVM, RF	SVM
(Alsubait and Alfageh, 2021)	Youtube	Arabic	Count Vectorizer, TF-IDF Vectorizer	F-Score: %78.6	MNB, CNB, LR	Count Vectorizer +LR
(Perera end Fernando, 2021)	Twitter	English	TF-IDF, Sentiment Analysis, Profanity	Accuracy: %75.17 Precision: %75 Recall: %75, F-Score: %75	SVM	SVM
(Balakrishnan et al., 2020)	Twitter	English	BigFive, Dark Triad,	Accuracy, AUC, F-Score, Kappa, RMSE	NB, RF, J48	RF, J48
(Dalvi et al., 2020)	Twitter	English	TF-IDF	Accuracy: %71.25 Precision: %71 Recall: %71, F-Score: %70	SVM, NB	SVM
(Muneer and Fati, 2020)	Twitter	English	TF-IDF, Word2Vec	Accuracy: %90.57, Precision: %95.18 Recall: %90.53,	LR, LGBM, SGD, MNB, RF,	LR

Study	Dataset		Method And Techniques			
	Platform	***Language***	***Feature Techniques***	***Evaluation Metrics***	***Classifiers***	***Best Classifier***
				F-Score: %92.8	AdaBoost, SVM	
(Talpur and O'Sullivan, 2020)	Twitter	English	POS tagging, PMI, Sentiment, Lexicon	Accuracy: %91.15, Kappa: %71.1 F-Score: %89.8	SVM, KNN, DT, RF, NB	RF
(Cuzcana and Ayma, 2020)	Twitter	Spanish	TF-IDF, N-Gram	Accuracy: %80 Precision: %81 Recall: %80, F-Score: %80	NB, MLR, SVM, RF	SVM
(Rosa et al., 2019)	Formspring.me, Bullying Traces V3.0	English	TF-IDF, Word Embeddings, Personality Trait Features, Textual Features, Sentiment Features, MRC Psycholinguistic Features	F-Score: %74 (Formspring.me) F-Score: %45 (Bullying Traces V3.0)	SVM	SVM
(Sahni and Raja, 2018)	Twitter	India, English	N-Gram, Sentiment Analysis	Accuracy: %98, Kappa: %98	NB, RF, J48	NB, RF, J48
(Altay and Alatas, 2018)	Formspring.me	English	TF-IDF, Word2Vec	Precision: %84.2, Recall: %83.2 F-Score: %83.2, ROC: %92	BLR, RF, Multi Layer Perceptron, J48, SVM	RF
(Hussain et al., 2018)	Formspring.me, Myspace	English	TF-IDF, N-Gram	F-Score:%95	SVM, SGD, RBF, LR	SGD
(Bozyigit et al., 2018)	Twitter	Turkish	BagofWords, TF-IDF and Information Gain	F-Score: %91	NB, SVM, RF, KNN, MNB, C4.5	SVM
(Abdullah-Al-Mamun and Akther, 2018)	Facebook, Twitter	Bangala	User-Based Features	Precision: %99, F-Score: %99 ROC: %71, Accuracy: %97.27	SVM, NB, J48, KNN	SVM
(Sintiha and Mostakim, 2018)	Twitter	English	-	Accuracy: %89.54	SVM, NB	SVM
(Ozel et al., 2017)	Instagram, Twitter	Turkish	Chi-Square and Information Gain	F-Score: %81	SVM, C4.5, MNB ve KNN	MNB
(Al-garadi et al., 2016)	Twitter	English	Chi-Square, Information Gain and Pearson Correlation	AUC: %94.3, F-Score: %93.6	NB, SVM, RF, KNN, RF+SMOTE (Proposed Method)	Proposed Method
(Huang et al., 2014)	Twitter	English	SMOTE, Information Gain	ROC: %75.5	J48, NB, SMO, Bagging and Dagging	Dagging
(Dadvar et al., 2013)	Yotube	English	content-based features, cyberbullying features, user-based features	Precision: %77, Recall: %55 F-Score: %64	SVM	SVM
(Dadvar et al., 2012)	Myspace	English	TF-IDF, Number of Foul Words, Profanity	Precision: %39, Recall: %6 F-Score: %15	SVM	SVM
(Reynolds et al., 2011)	Formspring.me	English	Lexicon-based features	Recall:78.5	C4.5, JRIP, IBK, SMO	IBK, C4.5
(Yin et al., 2009)	Kongregat, Slashdot, Myspace	English	TF-IDF, N-Grams	Precision: %35, Recall: %59 F-Score: %44	SVM	SVM

Table 3. Classical Neural Networks and Deep Learning Techniques Effectively Used in Cyberbullying

Study	Dataset		Method And Techniques			
	Platform	***Language***	***Feature Techniques***	***Evaluation Metrics***	***Classifiers***	***Best Classifier***
(Murshed et al., 2022)	Twitter	English	Word2Vec, TF-IDF, POS Tagging, Information Gain, SMOTE	Accuracy: %90.45, Precision: %89.52, Recall: %88.98, F-Score: %89.25, Specificity: %90.94	Bi-LSTM, RNN, SVM, MNB, RF and DEA-RNN (Proposed Method)	DEA_RNN
(Bharti et al., 2021)	Twitter	English	BagOfWords, Word Embedding using different GloVe	Accuracy: %92.6, Precision: %96.6, F-Score: %94.2	DT, NB, RF, XgBoost, SVM, LR, BLSTM	BLSTM
(On and Yeniterzi, 2020)	Twitter	Turkish	Word2Vec, FastText	F-Score: %93	NB, BLR, CNN	CNN
(Iwendi et al., 2020)	(Islam et al., 2021)	English	-	Accuracy: %82.18	BLSTM, GRU, LSTM, RNN	BLSTM
(Mounir et al., 2019)	Formspring.me	English	TF-IDF, N-Gram	Accuracy: %92.8, Precision: %92.4, Recall: %91.7, F-Score: %91.9	SVM, NN	NN
(Bozyigit et al., 2019	Twitter	Turkish	TF-IDF, N-Gram	F-Score:%91	ANN	ANN
(Banerjee et al., 2019)	Twitter	English	GloVe	Accuracy: %93,97	CNN	CNN
(Al-Ajlan and Ykhlef, 2018a)	Twitter	English	-	Accuracy: %95, Precision: %93, Recall: %73	CNN-CB (Proposed Method)	CNN-CB
(Curuk et al., 2018)	Formspring.me, Myspace	English	N-Gram, TF-IDF	F-Score: %95	Neural Network Based SVM, SGD, RBF, LR	SGD
(Haidar et al., 2018)	(Choi, Jeon and Kim, 2021)	Arabic	-	Accuracy: %91.17	FFNN	FFNN

4. DISCUSSIONS

In the study, many machine and deep learning approaches for the detection and prevention of cyberbullying were examined. Although the studies collected the data sets from similar platforms, it is expected that the success rates will be different since they work on different data sets.

One of the purposes of this review study is to present different techniques, algorithms and approaches used in the detection and prevention of cyberbullying. This study has contributed to the selection of classifiers with feature extraction and selection techniques to researchers who will work on the detection and prevention of cyberbullying in the future. In future studies, cyberbullying detection studies on photo, video and audio data in addition to textual data will be examined.

Since the data sets used in the examined studies have different properties, different preprocessing steps can be applied. In particular, the preprocessing stages and the methods applied in the detection of cyberbullying (feature extraction, feature selection, classifier etc.) may vary according to the features of the language used in the data set. The features of the language used in the studies should be well known and data labeling processes should be done by experts in this language.

5. CONCLUSION AND FUTURE WORK

Recently, the use of the internet and social networks has increased considerably. According to the report "We Are Social Digital 2022" report, 62.5% of the world's population is internet users and 58.4% are active social media users of the all users. These figures have reached double-digit growth rates, especially during the Covid-19 outbreak. Intensive use of the Internet and social networks may cause some harms along with its benefits. Social media has become a part of daily human life and there are serious mental, emotional and physical effects on people who are exposed to cyberbullying via social media. The long-term persistence of cyberbullying attacks on social networks also increases the severity of bullying. The detection and prevention of cyberbullies, who intentionally and continuously harm people, have become an important research topic. In this study, we considered the studies on the detection of cyberbullying under four main learning methods: supervised learning, lexicon-based learning, rule-based learning and mixed methods. Based on our review, machine learning algorithms is the mostly applied choice for researchers who aim to detect cyberbullying. The success of deep learning algorithms and the dynamic feature extraction steps motivate researchers to use them in detecting cyberbullying. When the first studies on the detection of cyberbullying were examined, it was seen that the success rates were quite low (Yin et al., 2009; Reynolds et al., 2011). However, these success rates have been increased with the use of different feature techniques and different algorithms. The applied feature extraction techniques have great influence on the success of the classifier and the running time. Therefore, fast and accurate determination of the feature extraction techniques to be applied on the dataset is important. The most commonly used feature techniques in cyberbullying detection are TF-IDF, Word2Vec, N-Gram, BagofWords, Chi-Square and Information Gain. Among them, TF-IDF, N-Gram and BagofWords are more successful than others in increasing the success of the classifier. Most of the studies in this field have been made on textual data and Twitter posts have been preferred as the data set, and English has been preferred as the data language. The most important reason for this is the easy availability of Twitter data sets and the widespread use of the English language. In addition, when the algorithms applied in this field are compared, it has been seen that SVM, NB and CNN algorithms give better results than others.

In next studies on detecting cyberbullying, it is aimed to identify the most effective attributes using local search techniques based on heuristic approaches for feature extraction. In addition to textual data, it is aimed to carry out cyberbullying detection studies on photographs, video and audio data. As a new perspective, it is planned to identify the communities that make cyberbullying attacks on social networks by using the methodology of complex network analysis together with cyberbullying detection in future studies.

REFERENCES

Abdhullah-Al-Mamun, Akhter S. (2018). Social media bullying detection using machine learning on Bangla text, 2018 10th International Conference on Electrical and Computer Engineering (ICECE), Dhaka, Bangladesh, pp. 385-388, doi:10.1109/ICECE.2018.8636797.

Al-Ajlan M. A., Ykhlef M. (2018a). Deep Learning Algorithm for Cyberbullying Detection, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, doi:http://dx.doi.org/10.14569/IJACSA.2018.090927.

Al-Ajlan M. A., Ykhlef M. (2018b). Optimized Twitter Cyberbullying Detection based on Deep Learning, 2018 21st Saudi Computer Society National Computer Conference (NCC), pp. 1-5, doi:10.1109/NCG.2018.8593146.

Alam K. S., Bhowmik S., Prosun P. R. K. (2021). Cyberbullying Detection: An Ensemble Based Machine Learning Approach, 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 710-715, doi:10.1109/ICICV50876.2021.9388499.

Al-garadi M. A., Varathan K. D., Ravana S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network, *Computers in Human Behavior*, pp. 433–443, doi:https://doi.org/10.1016/j.chb.2016.05.051.

Alsubait T., Alfageh D. (2021). Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments, *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 21, doi:https://doi.org/10.22937/IJCSNS.2021.21.1.1.

Altay E. V., Alatas B. (2018). Detection of Cyberbullying in Social Networks Using Machine Learning Methods, 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), Ankara, Turkey, pp. 87-91, doi:10.1109/IBIGDELFT.2018.8625321.

Balakrishnan V., Khan S., Arabnia H. (2020). Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning, Computers & Security, vol. 90, doi:101710.10.1016/j.cose.2019.101710.

Banerjee V., Telavane J., Gaikwad P., Vartak P. (2019). Detection of Cyberbullying Using Deep Neural Network, 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 604-607, doi:10.1109/ICACCS.2019.8728378.

Bauman S., Toomey B. R., Walker L. J. (2013). Associations among bullying, cyberbullying, and suicide in high school students. *Journal of Adolescence*, ISSN 0140-1971, pp. 341-350, doi:https://doi.org/10.1016/j.adolescence.2012.12.001.

Bharti S., Yadav A.K., Kumar M., Yadav D. (2021). Cyberbullying detection from tweets using deep learning, *Kybernetes*, doi:https://doi.org/10.1108/K-01-2021-0061.

Bozyigit A. (2020). A comprehensive cyberbullying dataset including social media features, doi:10.17632/pgfk7h4367.1.

Bozyigit A., Utku S., Nasiboglu E. (2018). Sanal Zorbalık İçeren Sosyal Medya Mesajlarının Tespiti, 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 281-281, doi:10.1109/UBMK.2018.8566529.

Bozyigit A., Utku S., Nasiboglu E. (2019). Cyberbullying Detection by Using Artificial Neural Network Models, 2019 4th International Conference on Computer Science and Engineering (UBMK), pp. 520-524, doi:10.1109/UBMK.2019.8907118.

Bozyigit A., Utku S., Nasibov E. (2021). Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, ISSN:0957-4174, doi:https://doi.org/10.1016/j.eswa.2021.115001.

Chatzakouy D., Kourtellisz N., Blackburnz J.,De Cristofaro] E., Stringhini] G., Vakali A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter, Proceedings of the 2017 ACM on Web Science Conference, doi:10.1145/3091478.3091487.

Cheng L., Silva Y. N., Hall D., Liu H. (2021). Session-Based Cyberbullying Detection: Problems and Challenges, *IEEE Internet Computing*, *25*, pp. 66-72, doi:10.1109/MIC.2020.3032930.

Choi Y., Jeon B., Kim H. (2021). Identification of key cyberbullies: A text mining and social network analysis approach, *Telematics and Informatics*,. 56, doi:https://doi.org/10.1016/j.tele.2020.101504.

Chudal R., Tiiri E., Klomek A. B. et al. (2021). Victimization by traditional bullying and cyberbullying and the combination of these among adolescents in 13 European and Asian countries, European Child & Adolescent Psychiatry, doi:https://doi.org/10.1007/s00787-021-01779-6.

Chun J., Lee J., Kim J., Lee S. (2020). An international systematic review of cyberbullying measurements, *Computers in Human Behavior*, *113*, ISSN 0747-5632, doi:https://doi.org/10.1016/j.chb.2020.106485.

Curuk E., Acı C., Essiz E. S. (2018). Performance Analysis of Artificial Neural Network Based Classfiers for Cyberbulling Detection, 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 1-5, doi:10.1109/UBMK.2018.8566566.

Cuzcano X. M., Ayma V. H. (2020). A Comparison of Classification Models to Detect Cyberbullying in the Peruvian Spanish Language on Twitter, *International Journal of Advanced Computer Science and Applications IJACSA*, vol. 11, doi:10.14569/IJACSA.2020.0111018.

Dadvar M., De Jong F. (2012). Cyberbullying detection: a step toward a safer internet yard, WWW '12 Companion: Proceedings of the 21st International Conference on World Wide Web, pp. 121–126, doi:https://doi.org/10.1145/2187980.2187995.

Dadvar M., De Jong F., Ordelman R. J. F., Trieschnigg R. B. (2012). Improved cyberbullying detection using gender information, Proceedings of the 12th Dutch-Belgian information retrieval workshop, pp.23-25.

Dadvar M., Trieschnigg D., Ordelman R., De Jong F. (2013). Improving cyberbullying detection with user context, Proceedings of the European conference on information retrieval 2013, pp. 693–696, doi:https://doi.org/10.1007/978-3-642-36973-5_62.

Dalvi R. R., Chavan S. B., Halbe A. (2020). Detecting A Twitter Cyberbullying Using Machine Learning, 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 297-301, doi:10.1109/ICICCS48265.2020.9120893.

Ghotra B., McIntosh S., Hassan A. E. (2017). A Large-Scale Study of the Impact of Feature Selection Techniques on Defect Classification Models, 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), pp. 146-157, doi:10.1109/MSR.2017.18.

Google Trends. (2021). Extracted on 10 November 2021 from https://trends.google.com/trends/explore?date=all&q=cyberbullying.

Haidar B., Chamoun M., Serhrouchni A. (2017). A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning, 2017 1st Cyber Security in Networking Conference (CSNet), Rio De Janeiro, Brazil, pp. 1-8, doi:10.25046/aj020634.

Haidar B., Chamoun M., Serhrouchni A. (2018). Arabic Cyberbullying Detection: Using Deep Learning, 2018 7th International Conference on Computer and Communication Engineering (ICCCE), pp. 284-289, doi:10.1109/ICCCE.2018.8539303.

Hemphill S. A., Kotevski A., Heerde J. A. (2015). Longitudinal associations between cyber-bullying perpetration and victimization and problem behavior and mental health problems in young Australians, *Int. J. Public Health*, vol. 60, no. 2, pp. 227–237, doi:10.1007/s00038-014-0644-9.

Hinduja S., Patchin J.W. (2008). Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization, *Deviant Behavior*, vol. 29, pp. 129-156, doi:10.1080/01639620701457816.

Huang Q., Singh V., Atrey P. (2014). Cyber Bullying Detection Using Social and Textual Analysis, SAM '14 2014, doi:10.1145/2661126.2661133.

Hussain M. G., Mahmud T. A., Akthar W. (2018). An Approach to Detect Abusive Bangla Text", 2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, pp. 1-5, doi:10.1109/CIET.2018.8660863.

Islam M. M., Uddin M. A., Islam L., Akter A., Sharmin S., Acharjee U. K. (2021). Cyberbullying Detection on Social Networks Using Machine Learning Approaches, 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1-6, doi: 10.1109/CSDE50874.2020.9411601.

Iwendi C., Srivastava G., Khan S., Maddikunta P. K. R. (2020). Cyberbullying detection solutions based on deep learning architectures, *Multimedia Systems 2020*, doi:https://doi.org/10.1007/s00530-020-00701-5.

Kargutkar S. M., Chitre V. (2020). A Study of Cyberbullying Detection Using Machine Learning Techniques, Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 734-739, doi:10.1109/ICCMC48092.2020.ICCMC-000137.

Kowalski R.M., Limber S.P. (2007). Electronic Bullying Among Middle School Students, *Journal of Adolescent Health*, pp. S22-S30, doi:10.1016/j.jadohealth.2007.08.017.

Manca S., Bocconi S., Gleason B. (2021). Think globally, act locally: A glocal approach to the development of social media literacy. *Computers & Education*, ISSN 0360-1315, doi:https://doi.org/10.1016/j.compedu.2020.104025.

Miller K. (2017). Cyberbullying and Its Consequences: How Cyberbullying Is Contorting the Minds of Victims and Bullies Alike, and the Law's Limited Available Redress, *Southern California Interdisciplinary Law Journal*, pp. 379-404.

Mounir J. H., Nashaat M., Ahmed M., Emad Z., Amer E., Mohammed A. (2019). Social Media Cyberbullying Detection using Machine Learning, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, pp. 703-707, doi:10.14569/IJACSA.2019.0100587.

Muneer A., Fati S.M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter, Future Internet, vol. 12, 187. doi:https://doi.org/10.3390/fi12110187.

Murshed B. A. H., Abawajy J., Mallappa S., Saif M. A. N., Al-Ariki H. D. E. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform, pp. 25857-25871, doi:10.1109/ACCESS.2022.3153675.

Nadali S., Murad M. A. A., Sharef N. M., Mustapha A., Shojaee S. (2013). A review of cyberbullying detection: An overview, 13th International Conference on Intellient Systems Design and Applications, pp. 325-330, doi:10.1109/ISDA.2013.6920758.

Nandhini S. B. and Sheeba I. J. (2015). Online Social Network Bullying Detection Using Intelligence Techniques. *Procedia Computer Science*, vol. 45, pp. 485–492, 2015, doi: 10.1016/j.procs.2015.03.085.

On E. P., Yeniterzi R. (2020). Cyberbullying Detection using Deep Learning and Word Embedding Analysis, 2020 28th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, doi:10.1109/SIU49456.2020.9302297.

Ozel S. A., Sarac E. (2017). Effects of Feature Extraction and Classification Methods on Cyberbully Detection, *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, pp. 190-200, doi:10.19113/sdufbed.20964.

Ozel S. A., Sarac E., Akdemir S., Aksu H. (2017) Detection of cyberbullying on social media messages in Turkish, 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, pp. 366-370, doi:10.1109/UBMK.2017.8093411.

Perera A., Fernando P. (2021). Accurate Cyberbullying Detection and Prevention on Social Media, Procedia Computer Science, vol. 181, pp. 605-611, doi:https://doi.org/10.1016/j.procs.2021.01.207.

Rao J., Wang H., Pang M., Yang J., Zhang J., Ye Y., Chen X., Wang S., Dong X. (2019). Cyberbullying perpetration and victimization among junior and senior high school students in Guangzhou, China, Injury Prevention, 25(1), pp. 13–19, doi:https://doi.org/10.1136/injuryprev-2016-042210.

Reynolds K., Kontostathis A., Edwards L. (2011). Using Machine Learning to Detect Cyberbullying, 10th International Conference on Machine Learning and Applications and Workshops, pp. 241-244, doi:10.1109/ICMLA.2011.152.

Roh Y., Heo G., Whang S. E. (2021). A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective, IEEE Transactions on Knowledge and Data Engineering, vol. 33, pp. 1328-1347, doi:10.1109/TKDE.2019.2946162.

Rosa H., Pereira N., Ribeiro R., Ferreira P.C., Carvalho J.P., Oliveira S., Coheur L., Paulino P., Veiga Simão A.M., Trancoso I. (2019). Automatic cyberbullying detection: A systematic review, *Computers in Human Behavior*, vol. 93, pp. 333–345, doi:10.1016/j.chb.2018.12.021.

Sahni A., Raja N. (2018). Analyzation and Detection of Cyberbullying: A Twitter Based Indian Case Study. In: Panda B., Sharma S., Roy N. (eds) Data Science and Analytics, REDSET 2017, Communications in Computer and Information Science, vol 799. Springer, Singapore. doi:https://doi.org/10.1007/978-981-10-8527-7_41.

Salawu S., He Y., Lumsden J. (2017). Approaches to automated detection of cyberbullying: A survey. Transactions on Affective Computing, IEEE Transactions on Affective Computing, vol. 11, no. 1, pp. 3-24, doi:https://doi.org/10.1109/taffc.2017.2761757.

Sintaha M., Mostakim M. (2018). An Empirical Study and Analysis of the Machine Learning Algorithms Used in Detecting Cyberbullying in Social Media, 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 1-6, doi:10.1109/ICCITECHN.2018.8631958.

Smith P. K., Mahdavi J., Carvalho M., Fisher S., Russell S., Tippett N. (2008). Cyberbullying: its nature and impact in secondary school pupils, *Journal of Child Psychology and Psychiatry*, pp. 376-85, doi:10.1111/j.1469-7610.2007.01846.x. PMID: 18363945.

Talpur B. A., O'Sullivan D. (2020). Cyberbullying severity detection: A machine learning approach, *PLoS ONE 15*(10): e0240924, doi:https://doi.org/10.1371/journal.pone.0240924.

Tiitle: "We Are Social". (2022). Extracted on 29 April 2022 from https://wearesocial.com/uk/blog/2022/01/digital-2022/.

Yin D., Xue Z., Hong L., Davison B. D., Kontostathis A., Edwards L. (2009). Detection of harassment on web 2.0. In Proceedings of the content analysis in the WEB (pp. 1–7).