Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 3 (No. 2), pp. 45-53, 2023
doi: 10.54569/aair.1206144
Published online: Oct 23, 2023

Creating a New Dataset for the Classification of Cyber Bullying

Çilem Koçak [1,*], Tuncay Yiğit [2], Mehmet Bilen [3]

[1] Isparta University of Applied Sciences, Yalvaç Vocational School of Technical Sciences, Isparta, Türkiye
[2] Süleyman Demirel University, Faculty of Engineering, Department of Computer Engineering, Isparta, Turkey
[3] Mehmet Akif Ersoy University, Golhisar School of Applied Science, Turkey

Abstract

Regardless of young or old, people have quickly stepped into the world of internet with today's communication technologies such as phones, tablets, computers and smart devices. As the place of the Internet in people's lives increases, social media platforms are diversifying and users want to take part in these platforms. With the increase in the number of social media users, some negativities are encountered. The most important problem encountered in social media platforms is cyber bullying. Although cyber bullying seems to be a daily dialogue between social media users or between groups, the situation of encountering is increasing day by day with the diversity of shared information, content and agenda social media environments. With the development of technology, it is necessary to develop a platform that detects bullying with artificial intelligence technologies. One of the biggest difficulties in text classification problems that we encounter during the development of these platforms is the need to train the artificial intelligence algorithm to be used with labeled data. In this study, 21 different people, including journalists, athletes, scientists, doctors, politicians, comedians, social media phenomena, and artists who actively use social media, were selected in order to create the necessary dataset for training the models to be developed to detect cyber bullying situations. The public messages (mentions) of these 21 people sent via Twitter were compiled. After filtering the repetitive and meaningless messages sent by bot accounts out of 10500 tweets compiled, the number of messages in the dataset decreased to 7706. The labeling process, which is necessary for the dataset to be used for training and testing purposes in classification processes, was carried out by three independent people who were given preliminary information about cyberbullying (1=Includes Cyber bullying, 0=Does not include Cyber bullying). The majority of the tags, which were read and assigned by 3 different people, were accepted as the final class of the relevant message. Afterwards, the dataset was preprocessed in accordance with the principles of natural language processing and made suitable for classification algorithms. The findings obtained after the classification processes performed with the basic classification algorithms are shared. When the findings are examined, it is understood that the data set created has the competence to be used in the detection and prevention of cyber bullying. In this context, it is predicted that training specially developed and optimized artificial intelligence algorithms with the relevant dataset for the detection of cyberbullying will greatly increase the success rate.

Keywords: *Cyber bullying; Twitter; artificial intelligence; text classification; data labeling.*

1. Introduction

With the development of telephone and computer technologies and the increase in the number of social media platforms, the likelihood of cyberbullying behaviors is also increasing. Victims of cyberbullying are threatened by users with electronic communication tools, often receive messages containing written insults, and face actions such as making someone look bad with a false identity. In this case, problems of mutual relations between the bully and the victim arise [1]. It is thought that these problems arise from the feeling of revenge caused by the deterioration of friendship and emotional relations between people, and the written disagreement between people who have different views and thoughts [2]. In the case of cyber bullying behavior, regardless of the means and environment in which the cyber bullying is carried out, it is desired to create a destructive result on the victim, to hurt, humiliate, humiliate and leave permanent traces on the victim, but the social relations of the victim are adversely affected. In addition, as a result of this, emotional, social and psychological damage occurs.

Since the platforms where cyberbullying behaviors are most common are social media, the increase in the number of users on these platforms directly increases the victimization [3]. In virtual environments;

- With its feature of hiding identity, it leads to the idea of having the right to say whatever they want to its victims.
- People can easily say things that they cannot say exactly what they want face to face, and they can

*Corresponding author
E-mail address: cilemkocak@isparta.edu.tr

Received: 17/Nov/2022; Received in revised form: 21/Mar/2023; Accepted: 22/Apr/2023.

isolate themselves by giving reactions that show that the other person does not approve or care about them.

- It allows victims to say what they want to say when they want and without censorship.
- Since people do not have any idea about the bully, they cannot express their thoughts and themselves clearly.
- It enables individuals to easily say their gender, social status, race and similar features that they cannot express in face-to-face communication.
- Individuals with an aggressive personality become more aggressive, causing them to express their personal style in an exaggerated way.

Looking at the rate of increase in the use of social media around the world, according to the January 2021 data of Statista [4], 63% of people around the world use Facebook, 61% use YouTube, 48% use WhatsApp, followed by Facebook Messenger, Instagram, Twitter and Snapchat are used.

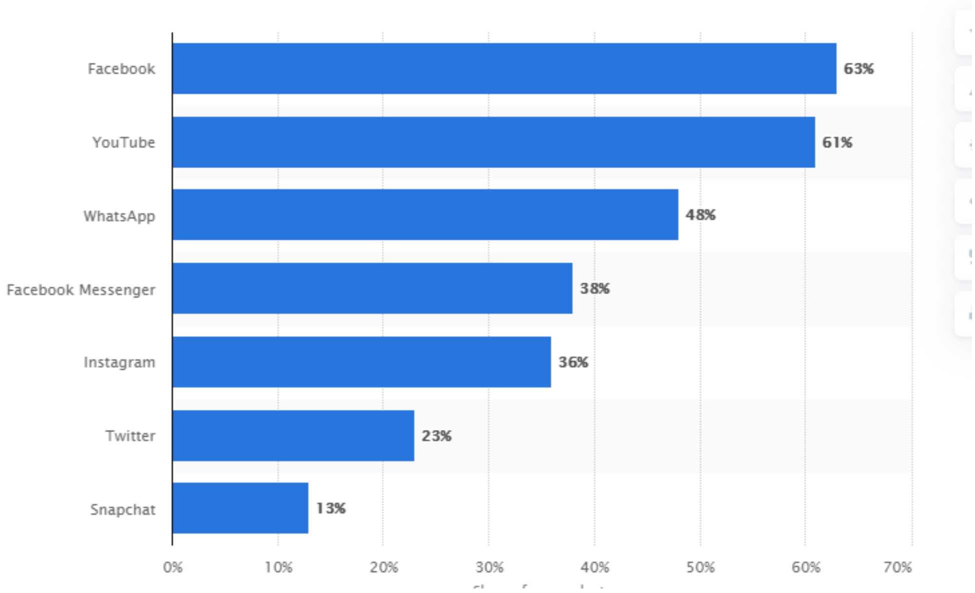

Figure 1. *We Are Social January 2021 Worldwide Social Media Usage Statistics*

According to the We Are Social 2021 report, looking at Turkey's internet, social media and mobile user statistics, 74% of Turkey's population is 62 million internet users, 64% of Turkey's population is 54 million social media users, 92% of Turkey's population is 77 million mobile users' forms. According to the We Are Social 2021 report, worldwide social media usage statistics are given in **Figure 1**.

As of 2021, the number of internet users in Turkey has approached 66 million with an increase of approximately 4 million within a year. This figure corresponds to a 6% increase in 1 year [5]. We see that there are 60 million social media users among the total population approaching 85 million. This means that 70.8% of the population is a social media user. In Turkey, 7 out of 10 people use social media and 9 out of 10 people use mobile devices. According to the results of the research, it is seen that those who are exposed to bullying behaviors experience different psychological disorders. These; mental illnesses such as sleep disorder, attention disorder, feeling of loneliness, depression. In some studies, it has been observed that the tendency to suicide increases in those who are exposed to cyber bullying [2].

Labeled datasets are needed to develop software that can detect cyber bullying behaviors. In this study, a new dataset was created in order to automatically detect cyber bullying behaviors and prevent the individual from being exposed to these behaviors.

In the continuation of the study, the concept of cyber bullying was examined and the studies carried out with machine learning in this field were mentioned in the literature. In the Material and Method heading, the steps followed in the dataset creation stages, the operations performed to make the dataset suitable for machine learning algorithms, and the classification steps are explained in detail. The results obtained in the research findings were interpreted and shared.

2. Cyber Bullying

Cyber bullying and types of bullying in the physical environment are similar to each other, although the environment in which the bullying takes place is different. It is seen that social media tools are used to exhibit cyber bullying behaviors. Bullies engage in cyber bullying in many ways. Common and classified types of cyber bullying;

- Cyber tracking: Keeping a person under constant surveillance in virtual environments,
- Slandering: Making false, harmful and rude statements about a person,
- Presenting Oneself as Someone Else: Impersonating an imaginary person or someone else by hiding their identity on the internet,
- Harassment: Sending offensive or sexually explicit messages to a person,
- Provoking: Encouraging a person for situations that he should not do,
- Wandering and Deception: Spreading embarrassing and private information about a person,
- Separation: Removing or not including a person from a group [1]

Žufic et al. (2017), Ayas and Horzum (2011), Karabatak et al. (2018), Arıcak(2011), Baker and Kavşut(2007), Arıcak, Siyahhan Uzunhasanoğlu, Sarıbeyoğlu, Çubuk, Yılmaz, and Memmedov (2008) tried to determine the information they have on cyber bullying by conducting surveys on the subject of cyber bullying, taking into account the age groups in different fields. carried out their studies. Within the scope of these studies, cyber bullying detection questionnaire and situation assessment questionnaire were used. As a result, it has an important place in the literature in terms of awareness of cyber bullying [6 - 11]. Hussain et al. (2018), Al-Mamun and Akhter (2018) used many artificial intelligence methods and techniques to detect cyber bullying data from different social media platforms [12,13]. The most used of these methods and techniques are; Dvm, Chi2, DVM-RFE, MRMR, C4.5 Decision Tree, k-nn classifiers, Maximum Entropy method, convolutional neural networks (Convolutional Neural Networks: CNN), bidirectional long short-term memories (Long Short-Term Memory: LSTM)) and Gated Recurrent Units (GRU), Naive Baeyes (NB), Random Forest (RF). In addition to these, there are different studies in the literature.

3. Material and Method

In this part of the study, firstly, how the dataset was obtained and the labeling process were explained in detail. Afterwards, the preprocessing performed on the dataset, the methods used to classify the text and the models used in this study for the classification of cyber bullying are presented respectively.

3.1. Dataset

The field of artificial intelligence contains many algorithms that learn from their experiences. A large amount of data is needed to obtain these experiences. In this study, it is aimed to collect data via Twitter social media application and to adapt this data to these algorithms in order to provide resources for artificial intelligence-based studies in the field of cyber bullying.

The tweets sent to the profiles of 21 different people who are famous in many different fields and actively use the Twitter social media application in the last 15 days were compiled using a program written in the Python programming language with the help of the API obtained by creating a Twitter developer account. These tweets must be tagged in order to be used in the training of artificial intelligence. Since the labeling process is done using human power, it is one of the processes that has the greatest impact on the development speed of artificial intelligence. In this study, tagging was carried out by students studying at the graduate level, who were given preliminary information about bullying and artificial intelligence. Each tweet was evaluated independently by three different people (1 = Contains bullying, 0 = No bullying) and the rating with the most votes was assigned to the class information of the relevant tweet. In this evaluation process, tweets that are thought to be sent from bot accounts, repeated and meaningless tweets were removed from the dataset with the initiative of the evaluators. At the end of the evaluation process, which started with 10500 tweets, it decreased to 7706.

3.2. Pretreatment

A large part of artificial intelligence algorithms depends on the number, type, size, etc. of data. is extremely sensitive. For this reason, preprocessing steps are of great importance for both adapting the data to artificial intelligence algorithms and increasing the success of these algorithms. Considering a traditional natural language processing preprocessing process and the characteristics of texts taken from social media, the following processing steps were followed in this study, respectively.

- Emoji cleaning
- Link, hashtag, mention information etc. cleaning up
- Conversion of uppercase letters to lowercase letters by considering Turkish characters
- Rooting each word (stem)
- Words, conjunctions, etc. that do not have a meaning on their own. removal from text (stopwords)

In order to extract the emojis in the sent tweets, the hexadecimal value of each character was obtained and compared with the intervals given in **Table 1**. Values within these ranges were removed from the text and the text was freed from emojis.

Table 1. *Unicode values of emojis*

Unicode Range	Emoji Type
0001F600-0001F64F	Feelings
0001F300-0001F5FF	symbols
0001F680-0001F6FF	Map and Logistics Symbols
0001F1E0-0001F1FF	Flags

In the queries written for cleaning links, hashtags and quotes, the characters and strings in **Table 2** were considered distinctive and those matching these characters were cleared from the data set. In addition, the different codes of uppercase and lowercase letters cause words that are actually the same to be perceived as different words by algorithms. For this reason, all capital letters were converted to lowercase letters, paying attention to Turkish characters.

Table 2. *Distinctive character strings*

Character Type	Type, Variety
#	hashtag
http://, https://, mailto:, www., .com	Links
@	Mention information
RT	Repost information

Since Turkish is an agglutinative language, natural language processing processes are more difficult than conventional (English, German, French, Spanish, etc.) languages. Unlike these languages, each word in Turkish can have more than one suffix. Each suffix causes the related word to be evaluated as another word. For this reason, it is necessary to purify the words from their suffixes. Although it is thought that the performance of artificial intelligence is negatively affected due to the elimination of attachments that contribute to the meaning in this process, this step should be carried out in order to minimize the operating cost of algorithms and to eliminate the incompatibility problems caused by excessive data complexity and data size. The “TurkishStemmer” library was used in the study developed with Python in order to separate the words into their roots [14]. The “turkishStopwords” list in NLTK [15], which is a frequently used natural language processing library, was used to extract words that are mentioned as stopwords in the literature and that do not have a meaning on their own. The words in the list are:

'acaba', 'ama', 'aslında', 'az', 'bazı', 'belki', 'biri', 'birkaç', 'birşey', 'biz', 'bu', 'çok', 'çünkü', 'da', 'daha', de', 'defa', 'diye', 'eğer', 'en', 'gibi', 'hem', 'hep', 'hepsi', 'her', 'hiç', 'için', 'ile', 'ise', 'kez', 'ki', kim', mı', 'mu', 'mü', 'nasıl', 'ne', 'neden', 'nerde', 'nerede', 'nereye', 'niçin', 'niye', 'o', 'sanki', 'şey', 'siz', şu', 'tüm', 've', 'veya', 'ya', 'yani'

3.3. Text Classification

The text classification problem can be summarized as determining or predicting which class the texts belong to according to the meaning they contain. It has become a critical work area with digital transformation. The fact that the amount of text that appears daily around the world has reached extremely large sizes has made it necessary to perform text classification automatically [16]. In order for the text to be analyzed with certain algorithms, it must be digitized. In this way, the data can be structured and the relationships between the data can be revealed [17]. When the studies in the literature in the field of text classification are examined, we come across three different methods for solving the problem;

- Rule Based Systems
- Machine Learning Based Systems
- Hybrid Systems

The rule-based approach is seen in the literature as a classification method in which the cause-effect relationship can be examined. With this method, the details of the classification process can be observed and additions can be made easily to improve the result [18]. Rule-based approaches consist of a set of if-not (IF-ELSE) rules [19]. Thanks to these rules, language-specific patterns are determined and it consists of structures that decide which class the relevant text belongs to. Jrip (Rajput, 2011), OneR (Buddhinath, 2006), ZeroR (Sayad, 2022) can be given as examples of these systems [20 - 22].

The biggest disadvantage of rule-based approaches is that rules must be created by linguistics experts. Contrary to rule-based approaches, machine learning algorithms used for text classification can learn from labeled data and create the relationships and rules between texts. In **Figure 2**, the flow diagram of traditional models that perform text classification with machine learning is given.

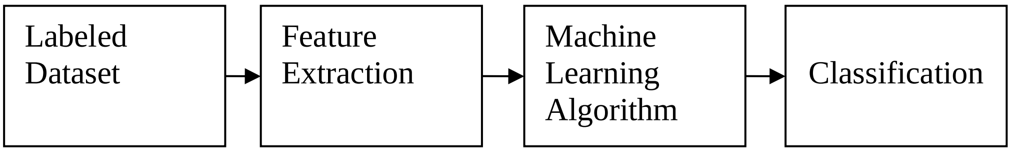

Figure 2. *Text classification model with machine learning*

In machine learning-based approaches, firstly, the labeled data is pre-processed to make it suitable for these algorithms and digitize it by feature extraction. Then, it is aimed to reveal the relationships between the data by using the new dataset obtained for the training of the algorithm. Finally, with the help of learned relations, a model emerges that can decide which class the new data belong to. Many machine learning algorithms such as artificial neural networks, support vector machine, Naive Bayes and Deep learning can be used by training for text classification purposes. However, due to the complex mathematical structures of these algorithms, the training phases take quite a long time. K Nearest Neighbors algorithm, which is an effective machine learning algorithm, was preferred in this study because of its simple structure and fast operation in order to determine whether the data set is classifiable or not.

Hybrid approaches, on the other hand, consist of a combination of a trained machine learning algorithm and the advantageous aspects of a rule-based approach to increase classification success.

3.4. Approach

The method followed to perform the training and classification processes with the machine learning algorithm after the compilation of the dataset is given in **Figure 3** as a flow chart.

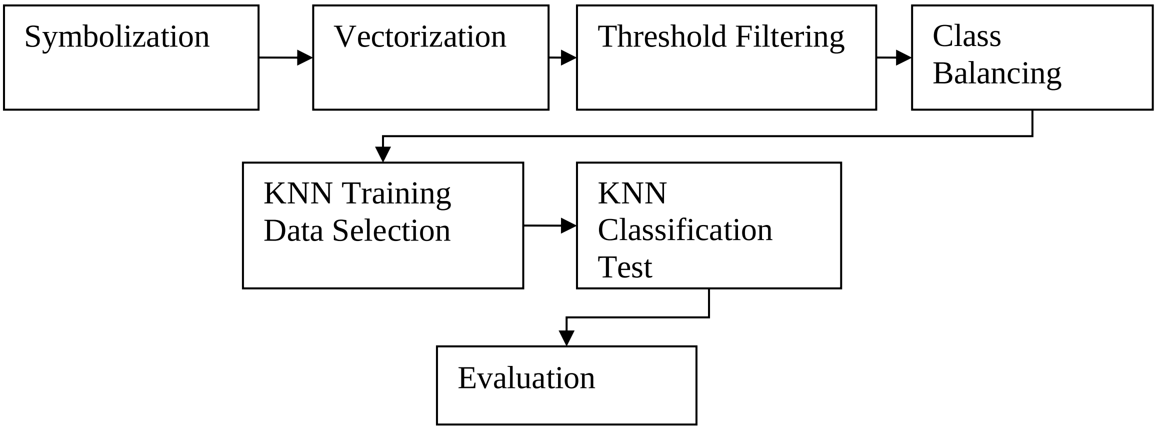

Figure 3. *Method followed for text classification*

Symbolization is simply the process of assigning a number to each word. In this way, a code transformation that algorithms can process is applied to the data. However, in this mathematical transformation, the numerical difference between a code of a text and another code has no meaning. For this reason, each sample in the dataset must be coded separately by representing which text it contains or does not contain with 0s and 1s. In this case, each sample has as many attributes as the number of unique texts and the size of the dataset grows and becomes more complex. In order to prevent this, words with a low number of repetitions were removed from the data set by selecting a certain threshold value. Just before the training and testing phase, a new data subset was created by randomly selecting the same number of samples from both classes in order to eliminate the uneven distribution of the classes in the dataset. 80% of the data obtained was used for training purposes to create the experience of the KNN algorithm. The remaining 20% was used for classification testing and the success of the model was tested. The results obtained were evaluated by sharing them under the title of research findings.

4. Results and Discussion

The tweets collected from the Twitter social media application are presented in **Figure 4**. When the figure is examined, a dataset with an unbalanced class distribution is seen. This imbalance is a situation that negatively affects the success of machine learning algorithms. Most of the machine learning algorithms are not sensitive to this imbalance and therefore they tend to predict in favor of the class with the large number of samples. To solve this problem, the samples in the 0 (No Cyberbullying) class were randomly eliminated and the number of samples in the two classes was equalized.

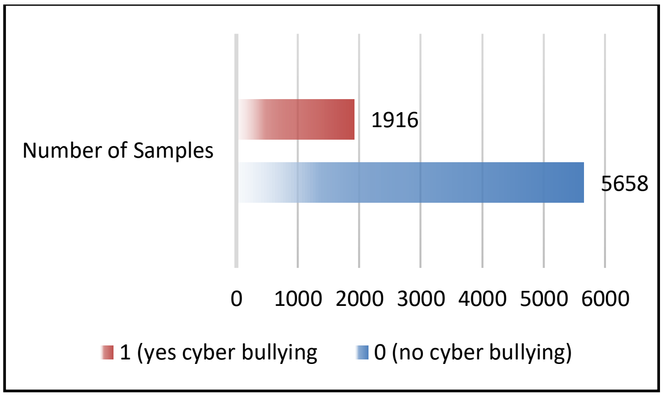

Figure 4. *Distribution of Sample Numbers by Classes*

In order to examine the characteristic features of the compiled tweets, the histogram of the number of words they contain is given in **Figure 5**. When the histogram is examined, it is seen that most of the tweets sent are under 10 words. It is expected that the number of words will be low due to the character limit set by the relevant social media application in the sent messages.

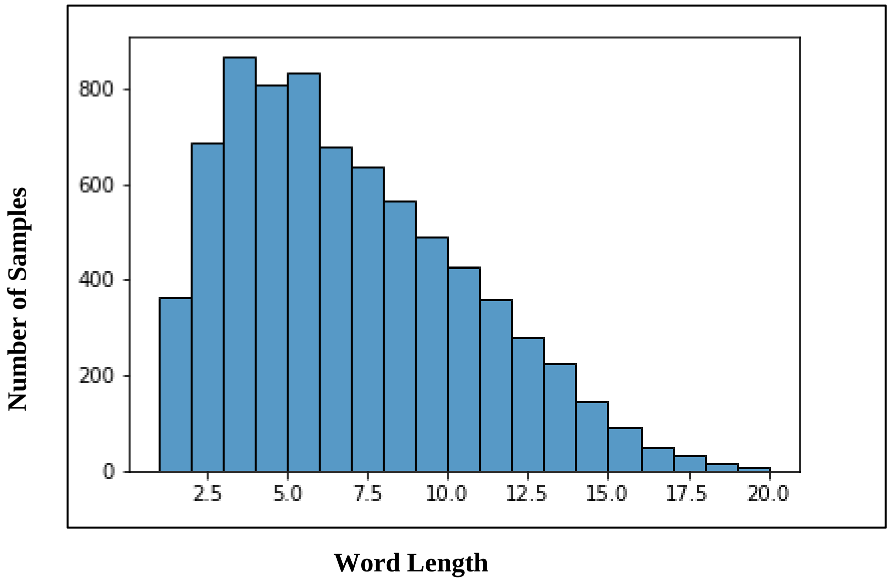

Figure 5. *Word Length Histogram*

The bi-grams in the dataset are sized depending on the repetition frequency and presented in Figure 6.

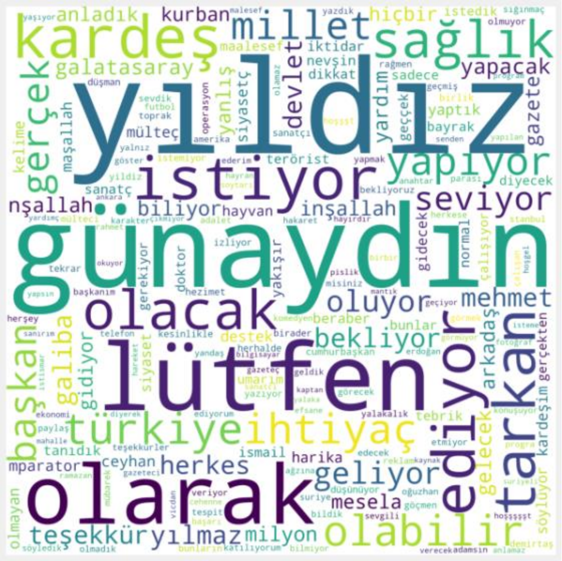

Figure 6. *Word Cloud (Frequency / Dimension)*

Classification analyzes after the preprocessing steps performed on the dataset and the selection processes performed to eliminate the imbalance were carried out by writing Python code on the Colaboratory Jupiter Notebook service provided by Google and on the virtual computer offered by Google.

As mentioned in the method section, the threshold value filtering method was used to reduce the complexity of the data size. Words with repetitions below a certain threshold were excluded from the dataset. All values between 1 and 250 were repeated to determine this threshold value (**Figure 7**). It is seen that the highest success in the classification processes performed by selecting different threshold values is obtained around 100 repetitions. When the operating time of the algorithm is examined (**Figure 8**), it is seen that the same threshold value has a positive contribution to the reduction of the time taken for analysis. Since the threshold value of 100 was confirmed in both graphs, this value was determined as the threshold value for the analyzes performed.

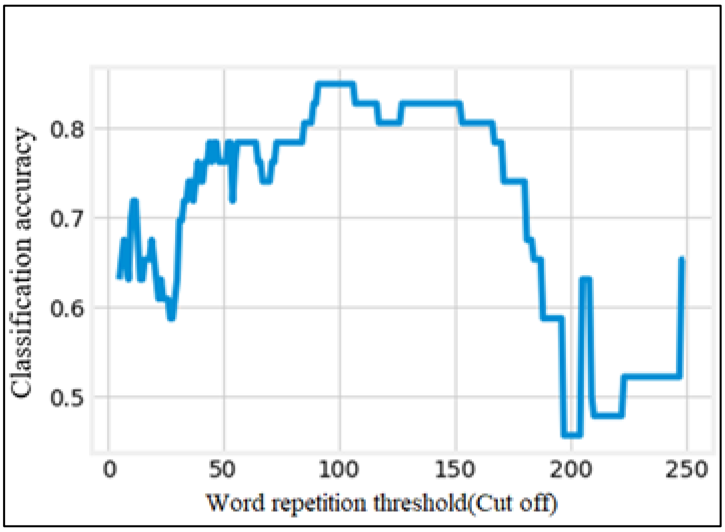

Figure 7. *Effect of Threshold Filtering on Classification Success*

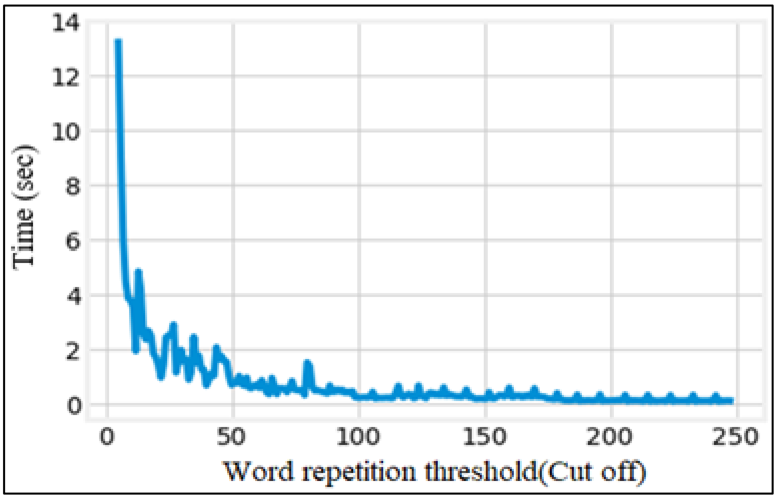

Figure 8. *Effect of Threshold Filtering on Algorithm Training and Classification Time*

KNN is a simple and fast but effective algorithm compared to other machine learning algorithms. The K value in the algorithm is one of the most important parameters that determine the success performance of this algorithm. Since there is no linear method to determine this parameter, which differs according to each data set, it is aimed to determine the most appropriate K parameter by iterating with different K values, as in the threshold value determination method. The graph obtained as a result of iteration is presented in **Figure 9**. When the graph is examined, it is seen that the value of 3 is the most appropriate K parameter in the classification processes to be performed with KNN on the data set prepared in this study.

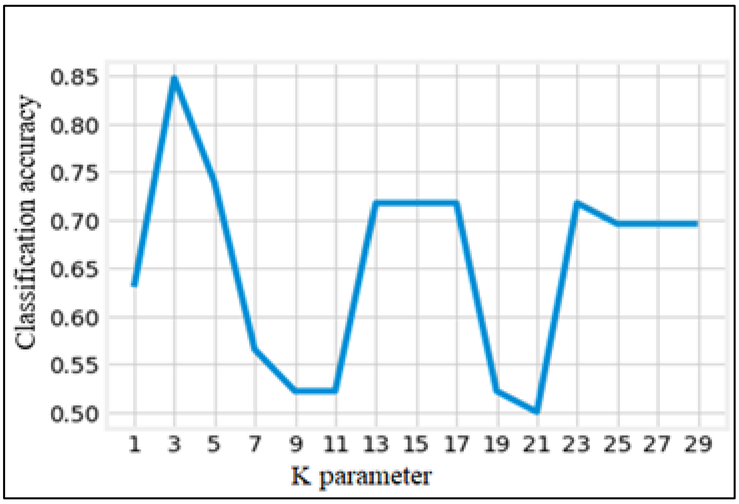

Figure 9. *Effect of K Parameter on Classification Result*

The findings obtained from the classification test processes performed after the determination of the most appropriate word repeat threshold value and the K parameter are presented in **Table 3**. The table shows that the prepared dataset can be successfully classified even with an algorithm that requires little training, such as KNN. It is seen that the prepared dataset is distinguishable and can identify the tweets containing bullying with a success rate of 85%.

Table 3. *Classification Performance Values*

	Precision	Recall	F1-Puanı
0 – No Bullying	0.79	0.9	0.87
1 – There Is Bullying	0.94	0.73	0.82
Percentage of Success			85
Macro Average	0.87	0.84	0.84
Micro Average	0.86	0.85	0.85

5. Conclusion and Suggestions

In this study, a new dataset was created to be used in machine learning algorithms to be developed to detect cyber bullying behaviors. As a result of the improvement processes and analyzes performed on the data set, it is seen that the data set can be classified at a rate of 85% with the KNN algorithm. It is thought that this rate will increase greatly with the evaluation of the dataset by algorithms with long training time such as deep learning.

Declaration of interest

The authors declare that there is no conflict of interest. It was presented as a summary at the ICAIAME 2022 conference.

References

[1] Gezgin, D. M., & Çuhadar, C. “Bilgisayar ve öğretim teknolojileri eğitimi bölümü öğrencilerinin siber zorbalığa ilişkin duyarlılık düzeylerinin incelenmesi”, *Eğitim Bilimleri Araştırmaları Dergisi*, *2*(2) (2012), 93-104.

[2] Özdemir, M., & Akar, F. “Lise Öğrencilerinin Siber-Zorbalığa İlişkin Görüşlerinin Bazı Değişkenler Bakımından İncelenmesi”, Kuram ve Uygulamada Eğitim Yönetimi, 4(4) (2011), 605-626.

[3] Eroğlu, Y., Güler, N. “Koşullu Öz-Değer, Riskli İnternet Davranışları ve Siber Zorbalık/Mağduriyet Arasındaki İlişkinin İncelenmesi”, Sakarya University Journal Of Education, 5(3) (2015), 118-129.

[4] Global social media usage report 2021, https://www.statista.com/ (accessed: Apr 10, 2022).

[5] Turkey Internet, social media and Mobile User Statistics According to We Are Social 2020-2021 Report Https://Wearesocial.Com/ (accessed: Jun 15 2022).

[6] Žufić, T. Žajgar, S. Prkić, “Children Online Safety”, 2017 40th International Convention On Information And Communication Technology, Electronics And Microelectronics (MIPRO), 22-26 May 2017, Opatija, Croatia

[7] Ayas, T., & Horzum, M. B. (2011). Exploring The Teachers' Cyber Bullying Perception In Terms Of Various Variables. International Online Journal of Educational Sciences, 3(2).

[8] S. Karabatak, A. Namlı, M. Karabatak, “Perceptions of High School Students Regarding Cyberbullying and Precautions on Coping With Cyberbullying”, 2018 6th International Symposium On Digital Forensic And Security (ISDFS), 22-25 March 2018, Antalya, Turkey.

[9] Arıcak, O. T. “Siber Zorbalık: Gençlerimizi Bekleyen Yeni Tehlike”, Kariyer Penceresi, 2(6) (2011), 10-12

[10] Erdur-Baker, Ö. and Kavşut, F. “Akran Zorbalığının Yeni Yüzü: Siber Zorbalık”, Eurasian Journal of Educational Research (EJER), 27(2007), pp, 31-42.

[11] Aricak, T., Siyahhan, S., Uzunhasanoglu, A., Saribeyoglu, S., Ciplak, S., Yilmaz, N., & Memmedov, C. Cyberbullying Among Turkish Adolescents. Cyberpsychology & Behavior, 11(3) (2008), 253-261.

[12] M. G. Hussain, T. Al Mahmud, W. Akthar, “An Approach To Detect Abusive Bangla Text”, International Conference On Innovation İn Engineering And Technology (ICIET), 27-29 December 2018.

[13] Al-Mamun, S. Akhter, “Social Media Bullying Detection Using Machine Learning On Bangla Text”, 10th International Conference On Electrical And Computer Engineering, 20-22 December 2018, Dhaka, Bangladesh

[14] Turkishstemmer. Https://Github.Com/Otuncelli/Turkish-Stemmer-Python (accessed: Jun 13, 2022).

[15] NLTK: Https://Www.Nltk.Org/ (accessed: Jun 10, 2022).

[16] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text Classification Using Machine Learning Techniques. WSEAS Transactions On Computers, 4(8), 966-974.

[17] Wilkinson, A. W. Literature Review on Advance Directives. U.S. Department of Health and Human Services. Washington: RAND Corporation, 2007.

[18] Abuaid, A. M., & Mishra, A. (2010). A Rule-Based Approach to Embedding Techniques for Text Document Classification. Applied Science, 10(11), 4009.

[19] Ross, T. J. (2005). Fuzzy Logic with Engineering Applications. West Sussex, United Kingdom: John Wiley & Sons.

[20] Rajput, A. A. (2011). J48 And JRIP Rules For E-Governance Data. International Journal of Computer Science and Security, 5(2), 201.

[21] Buddhinath, G. D. (2006). A Simple Enhancement To One Rule Classification. Melbourne, Australia: Department of Computer Science & Software Engineering University Of Melbourne.

[22] Sayad, S. (2022). Zeror, Saedsayad: Https://Www.Saedsayad.Com/Zeror.Htm (accessed: Jun 10, 2022).