Düzce University
Journal of Science & Technology

Research Article

Detecting Turkish Cyberbullying Tweets Using Machine Learning

Yavuz Selim BALCIOĞLU [a,*]

[a] *Department of Management Information System, Faculty of Social Science, Gebze Tehcnical University, Kocaeli, TURKEY*
* *Corresponding author's e-mail address: ysbalcioglu@gtu.edu.tr*

DOI: 10.29130/dubited.1379657

ABSTRACT

Cyberbullying is a form of crime where individuals are subjected to online hate speech and harassment, and its prevalence has increased with the growth of social media. There is a noticeable gap in the current literature, especially for cyberbullying detection in languages other than English. This study proposes a method for automatic cyberbullying detection in Turkish tweets. The proposed model incorporates the Support Vector Machine and Random Forest classification algorithms. The model has been trained on labeled real-world data sourced from Twitter. To address the characteristics of the Turkish language, a natural language processing tool called Zemberek-NLP has been used. This tool captures the nuances of the language, enhancing the accuracy of the detection model. This research aims to contribute to the fight against cyberbullying by presenting an innovative approach to detecting it in Turkish.

Keywords: *Cyberbullying detection, Turkish social media, Machine learning, Support Vector Machine (SVM), Random Forest (RF) classifier*

Makine Öğrenimi Kullanarak Türkçe Siber Zorbalık Tweetlerini Tespit Etme

ÖZ

Siber zorbalık, çevrimiçi nefret söylemi ve tacizle bireylerin maruz kaldığı bir suç biçimidir ve sosyal medyanın büyümesiyle yaygınlık kazanmıştır. Mevcut literatürde, özellikle Türkçe dışındaki dillerde siber zorbalık tespiti için belirgin bir eksiklik bulunmaktadır. Bu çalışma, Türkçe tweet'lerde otomatik siber zorbalık tespiti için bir yöntem önermektedir. Önerilen model, Destek Vektör Makinesi ve Rastgele Orman sınıflandırma algoritmalarını içerir. Model, Twitter'dan alınan etiketli gerçek dünya verisiyle eğitilmiştir. Türk, dilinin, özelliklerini ele almak için Zemberek-NLP adlı bir doğal dil işleme aracı kullanılmıştır. Bu, araç, dilin, nüanslarını, ele alarak, tespit modelinin doğruluğunu artırır. Bu çalışma, Türkçe'deki siber zorbalık tespiti için yenilikçi bir yaklaşım sunarak, siber zorbalıkla mücadeleye katkıda bulunmayı hedeflemektedir.

Anahtar Kelimeler: *Siber zorbalık tespiti, Türkçe sosyal medya, Makine öğrenimi, Destek Vektör Makinesi (SVM), Random Forest (RF) sınıflandırıcı*

Received: 22/10/2023, Revised: 26/12/2023, Accepted: 21/02/2024

I. INTRODUCTION

In recent years, the digital landscape has undergone a striking transformation with the evolution of the internet and social media platforms into powerful venues for education, discourse, and idea exchange. Twitter, one of the leading social network platforms [1], provides a dynamic environment that allows users to broadcast both positive and negative thoughts beyond geographical and temporal boundaries. These virtual communities, particularly appealing to the younger demographic [2], have become an integral part of our increasingly digital world. However, the anonymity offered by these platforms, where users typically prefer pseudonyms over real names, contributes to an uptick in online violations, including cyberbullying [3]. These clandestine activities pose significant challenges for monitoring and regulation. Cyberbullying, defined as any aggressive act directed at an individual through online media, is a critical ethical concern on the internet. The number of individuals [4], especially youth, falling victim to cyberbullying, is alarmingly high. Cyberbullying has been the focal point of numerous studies aiming to ascertain its prevalence, and the results consistently underscore it as a persistent issue among today's youth, with the number of victims showing an increasing trend [5]. In response to these challenges, researchers have pioneered a range of mechanisms for the detection of cyberbullying, aiming to enhance surveillance and foster preventive measures.

Research in the field of cyberbullying detection has seen a significant surge in recent years. However, regarding addressing this issue in languages other than English, particularly in Turkish, there exists a notable gap. The Turkish language possesses a rich morphological structure that adds a layer of complexity to the task of detecting cyberbullying. This intricacy, combined with the dearth of specific research in this field, presents a significant challenge in devising an effective detection mechanism that takes into account the Turkish cultural and linguistic context. In light of cross-cultural and linguistic variances, users and their interactions present multifaceted challenges. Solutions formulated for other linguistic landscapes are not readily transferrable to Turkish contexts, given the distinctive cultural subtleties and idiosyncratic expressions inherent to the language. Some phrases or terms, which might seem harmless or even mundane in certain cultures, can carry aggressive connotations in the Turkish context. Despite these challenges, recent advancements in the automatic detection of cyberbullying have resulted in notable developments in classifying cases of cyberbullying, especially in the English language [6]. However, research geared towards the application of machine learning for Turkish cyberbullying detection on social networks remains inadequate.

This paper, while acknowledging the aforementioned challenges, identifies a significant gap in the current literature: a lack of research aimed at detecting cyberbullying in non-English social media content, particularly in Turkish. We propose an automatic cyberbullying detection method for Turkish tweets using machine learning techniques and the Zemberek-NLP tool, aiming to overcome the complexities of the Turkish language. By focusing on this niche yet critical research area, this paper aspires to contribute to the broader fight against cyberbullying striving to create safer online spaces for users beyond linguistic and cultural boundaries.

As we proceed, the article is structured into several key sections to provide a comprehensive understanding of our research. Following this introduction, we explore into the Literature Review, where we explore existing studies and methodologies pertinent to cyberbullying detection, particularly focusing on the unique challenges posed by the Turkish language. The subsequent section, Proposed Framework and Approach, outlines the methodology we employed in our study. Here, we detail our approach to data collection, preprocessing, and the specific machine learning algorithms used, namely the Support Vector Machine (SVM) and Random Forest (RF) classifiers. In the Analysis Results and Discussion section, we present the outcomes of our experiments, comparing the effectiveness of different classifiers and preprocessing techniques. This section is crucial for understanding the efficacy of our proposed model in the context of Turkish cyberbullying detection. Finally, the Conclusions section encapsulates our key findings, highlights the implications of our research, and suggests directions for future work. This includes potential applications of our model in broader social media contexts and the exploration of deep learning techniques for enhanced cyberbullying detection.

II. BACKGROUND

In this section, we will explore the fundamentals of cyberbullying, its impact, and the role of Machine Learning and Natural Language Processing (NLP) in addressing this issue.

A. CYBERBULLYING: DEFINITION AND TYPES

As a complex and multifaceted phenomenon, cyberbullying lacks a universally accepted definition. However, it has been explored from various angles in the literature, each providing a unique interpretation. In a general sense, cyberbullying is conceptualized as a mode of harassment mediated through Information and Communication Technologies (ICT). This encompasses a spectrum that spans textual data, messaging platforms, and an array of social media channels. An alternative widely accepted delineation of cyberbullying describes it as "a deliberately recurring act executed by an individual or collective, leveraging electronic modalities, directed at a victim who faces challenges in mounting a sustained defense.

Cyberbullying manifests itself in multiple forms, including:

- Threats: The perpetrator sends intimidating messages to instill fear in the victim.
- Persistent Harassment: The aggressor persistently transmits identical remarks or incendiary comments, or repeatedly engages the “Enter” key, thereby obstructing the victim's ability to partake in the discourse.
- Masqueraded Attack: The bully assumes another person's identity, creating an illusion that the intimidation isn't direct.
- Mass Attack: A perpetrator sends disparaging or rude messages to one or more victims in an online group via email or electronic messages.
- Trolling: The perpetrator intentionally posts distasteful comments to incite negative discussion or emotions.
- Harassment: The perpetrator persistently sends aggressive messages to users.
- Denigration: Often referred to as 'Dissing', this entails a perpetrator propagating unfounded rumors or misleading information concerning an individual, aiming to malign their standing or interpersonal relations.
- Social Exposure: The aggressor reveals confidential or humiliating details pertaining to the victim on widely accessible social media platforms.
- Exclusion: The intentional exclusion of an individual from a social community, a form of bullying commonly seen among teenagers and adolescents.

As technology continues to progress and proliferate, it brings forth a series of ethical dilemmas. Social networks, with their vast user base, have become indispensable in many societies, including Turkey. Recent research in 2023 has revealed intriguing trends. The percentage of young Turkish users on social media has seen a steady increase from 35% in 2015 to a remarkable 85% in 2023. As of 2023, Turkey ranks among the top ten countries globally in terms of active Twitter users. While this hyper-connectivity offers countless opportunities for communication and learning, it also harbors various risks, one of which is cyberbullying. Given the potentially devastating consequences for its victims, this issue has grown into a significant societal concern [7]. A study conducted in Turkey between 2022-2023 revealed that individuals aged 10 to 19 who experienced cyberbullying or online harassment were 60% more likely to have suicidal thoughts. Additionally, research by the Cyberbullying Research Group in Turkish schools indicated that over 42% of students aged 12-17 skipped school due to their experiences with cyberbullying [8]. The adverse effects of cyberbullying extend beyond student performance, impacting the victim's mental health and self-esteem. Cyberbullying statistics underscore the severity of this issue and the urgent need for effective countermeasures. The repercussions on public health are profound, with 46% of cyberbullying victims exhibiting social anxiety, 40% showing signs of depression, and 30% having suicidal thoughts [9]. These figures underscore the pressing need for robust solutions to address the cyberbullying problem on Turkish social media platforms.

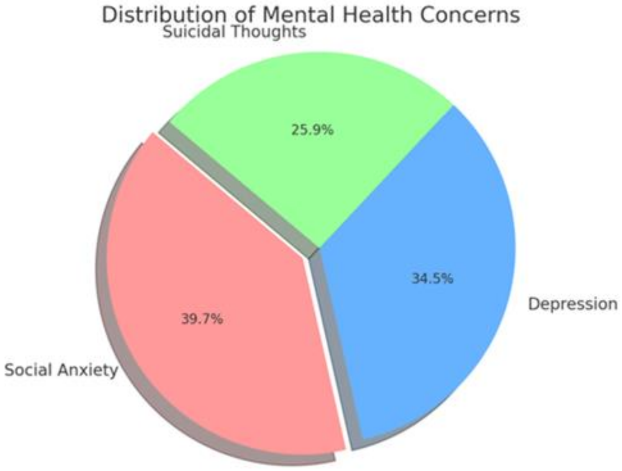

Figure 1. *Distribution of mental health concerns*

B. MACHINE LEARNING

As a subset of Artificial Intelligence (AI), Machine Learning (ML) empowers systems with the ability to learn and improve from past experiences [10] without the need for explicit programming, thereby enhancing automation processes. Essentially, machine learning allows a system to learn autonomously based on a training data set, facilitating decisions. This capability is especially useful for tasks like cyberbullying detection, which can be too intricate or time-consuming for humans. Machine Learning is broadly divided into two approaches: supervised and unsupervised learning [11]. Supervised learning algorithms utilize a labeled training dataset to construct a predictive model. This model can later be used to predict class labels for unlabeled data. Classification methodologies exhibiting commendable efficacy and precision encompass Decision Trees [12], Naïve Bayes (NB), K-Nearest Neighbors, Support Vector Machine (SVM), and Random Forest (RF). Within the scope of our investigation, the SVM and RF algorithms, employed as binary classifiers, operate under the presumption that data instances are delineable with discernible demarcations. They attempt to determine the optimal hyperplane (or multiple hyperplanes in the case of RF) that maximizes the margin between classes. On the other hand, unsupervised learning algorithms use unlabeled training data. In the absence of predetermined classifications, the system endeavors to discern inherent patterns within the data and cluster analogous data points in proximity. While our study does not directly use unsupervised learning, it seeks to highlight its potential use in exploratory analysis or in the preprocessing stages of the cyberbullying detection process.

C. MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING IN CYBERBULLYING DETECTION

Machine Learning (ML) and Natural Language Processing (NLP) have emerged as crucial tools in detecting and mitigating cyberbullying. Machine Learning algorithms like Support Vector Machine (SVM) and Random Forest (RF) allow for automated detection by being trained on annotated datasets to classify the content as cyberbullying or not. Meanwhile, Natural Language Processing tools for Turkish, such as Zemberek-NLP, are used to analyze and understand the linguistic nuances of the content. NLP techniques like tokenization, part-of-speech tagging, and named entity recognition assist in extracting meaningful features from text data, which can later be used to train machine learning algorithms.

III. LITERATURE REVIEW

In this section, we present a comprehensive review of the literature on cyberbullying detection, categorizing the studies into two primary areas: problem definition and technical methodologies.

A. PROBLEM DEFINITION IN CYBERBULLYING RESEARCH

Recent research has emphasized automatic cyberbullying detection, leveraging users' psychological characteristics [13] presents a study that developed a cyberbullying detection framework using Twitter content. The authors proposed leveraging a pointwise mutual information technique to generate features, which were then used in a supervised machine learning solution for both detecting cyberbullying and categorizing its severity. The study applied various features, including Embedding, Sentiment, and Lexicon, along with PMI-semantic orientation, to algorithms such as Naïve Bayes, KNN, Decision Tree, Random Forest, and Support Vector Machine. The results were promising in both multi-class and binary settings, showing high classifier accuracy and f-measure metrics. This study is significant as it highlights the feasibility of using advanced feature generation techniques in the effective detection of cyberbullying behavior and its severity on social networks, [14] critically examines the existing research on cyberbullying, noting the challenges posed by inconsistent findings and exaggerated prevalence claims. The authors emphasize the importance of reaching a consensus on defining cyberbullying as a scientific concept, distinct from general cyberaggression or cyberharassment. They advocate for considering cyberbullying within the broader 'bullying context,' which would facilitate clearer and more focused research outcomes. The recommendation to categorize cyberbullying as a specific form of bullying, akin to verbal, physical, and indirect/relational bullying, is pivotal. This perspective is crucial for developing targeted detection and intervention strategies, as it delineates cyberbullying from other forms of online aggression, thereby refining the focus of machine learning algorithms used for detection. Studies have explored aspects like personality, emotion, and sentiment on Twitter, using models such as the Big Five and the Dark Triad. In [15] involves a survey of 2052 school children, shedding light on the prevalence of cyberbullying and the discrepancies between direct and indirect measurement methods. Key findings indicate that younger children who engage in cyberbullying are often involved in traditional bullying, either as perpetrators or victims. The study also reveals that victims of cyberbullying tend to depend more on the internet, perceive themselves as less popular, engage in riskier online behaviors, and are frequently involved in cyberbullying as both bystanders and perpetrators. These insights are vital for understanding the multifaceted nature of cyberbullying and its overlap with traditional bullying. The study's implications for future research and prevention strategies provide a foundational basis for developing more effective cyberbullying detection and intervention programs, particularly those targeting school-aged children. Natural Language Processing (NLP) techniques are applied to correlate linguistic characteristics of tweets with these psychological frameworks, indicating that incorporating psychological dimensions enhances the accuracy of detection algorithms.

B. TECHNICAL METHODOLOGIES IN CYBERBULLYING DETECTION

In the technical realm, various machine learning classifiers have been employed for cyberbullying detection. A study focusing on the use of machine learning in social media, [16] addresses the growing issue of cyberbullying in the context of the increased use of social media. It highlights how social networks provide a fertile ground for bullying behavior and underscores the necessity of detecting and preventing cyberbullying due to its adverse effects on victims. The paper proposes a supervised machine learning approach, utilizing several classifiers to train and recognize bullying actions. The evaluation of their approach on a cyberbullying dataset demonstrates the superior performance of Neural Networks, achieving an accuracy of 92.8%, and compares favorably to SVM, which achieves 90.3% accuracy.

These results are particularly noteworthy as they indicate that Neural Networks are more effective than other classifiers for detecting cyberbullying patterns, which aligns with the increasing focus on machine learning as a vital tool in combating online harassment. Another significant contribution is a comprehensive review of cyberbullying prediction models in the context of social media platforms [17]. The authors explore the transformation of social interactions from geospatially bound communication to the expansive domain of online platforms. This shift has resulted in new forms of online aggression and violence, including cyberbullying. The paper highlights the importance of constructing prediction models to combat aggressive behaviors on social media. A key focus of this research is the review of cyberbullying prediction models, identifying the main challenges in developing these models for social media contexts. The study emphasizes the methodology involved in cyberbullying detection, covering aspects such as data collection, feature engineering, and the application of machine learning algorithms for predicting cyberbullying behavior. The paper concludes by presenting the issues and challenges in this field, thereby offering new research directions for scholars to investigate further. This study traces the evolution of social interactions and the application of various machine learning techniques for cyberbullying behavior prediction, emphasizing the importance of accuracy, precision, recall, and f-measure as evaluation metrics. The study focusing on sarcasm in cyberbullying, [18] addresses the challenge of cyberbullying in the context of the widespread use of social media platforms. It underscores that, while various strategies have been proposed to combat cyberbullying, the aspect of sarcasm remains relatively unexplored. This research aims to fill this gap by proposing an approach that not only detects cyberbullying but also considers the element of sarcasm. Such an approach is crucial because sarcasm adds a layer of complexity to the detection process, often masking the intent of the message. The study's findings indicate that the Support Vector Machine (SVM) classifier outperforms other classifiers in this context. This insight is particularly valuable for developing more sophisticated cyberbullying detection tools that can accurately interpret and categorize sarcastic content, which is often used in harmful ways online. Furthermore, the study introducing the Participant-Vocabulary Consistency (PVC) method, [19] tackles the issue of harassment and cyberbullying on social media, a problem that has escalated in scale and severity. The study proposes a unique machine learning method that infers user roles in harassment-based bullying while also identifying new vocabulary indicators. This method, which requires only weak supervision, uses a seed vocabulary provided by experts and then applies a large, unlabeled corpus of social media interactions to further extract bullying roles and language indicators. The key aspect of this approach is the Participant-Vocabulary Consistency (PVC) model, which estimates the bullying nature of interactions based on participant behavior and language use. The effectiveness of PVC in cyberbullying detection has been demonstrated through evaluations on three different social media datasets. This approach is significant as it offers a comprehensive way to understand both the social structure of bullying and the linguistic cues that accompany such behaviors, thereby enhancing the detection and analysis of cyberbullying incidents. The prevalence of cyberbullying in Turkey [20] has underlined the need for effective detection mechanisms. Studies have shown promising results using Linear SVM models with text vectorization methods like CountVectorizer and Tf-Idf Vectorizer [21]. These findings indicate that machine learning is pivotal in addressing cyberbullying, especially in languages other than English. Additionally, recent advancements in Turkish cyberbullying detection include the development of eight different artificial neural network models, such as ANN-2, which demonstrated a 91% F-measure score in identifying cyberbullying in 3000 Turkish tweets, outperforming several machine learning classifiers from previous studies [22]. Collectively, these studies provide a nuanced understanding of cyberbullying detection, highlighting the advancements in machine learning techniques and the importance of considering various linguistic and social factors.

Collectively, these studies indicate that, while significant advancements have been made in the field of cyberbullying detection [23], there remains a significant room for improvement, especially when considering the content in Turkish. Hence, in this scholarly composition, our objective is to traverse the intricacies inherent to the Turkish language utilizing machine learning methodologies in conjunction with the Zemberek-NLP instrument. Our ultimate goal is to enhance the efficacy of cyberbullying detection within the Turkish digital landscape.

The primary aim of our research is to develop an effective method for detecting cyberbullying in Turkish tweets using machine learning techniques. This objective stems from the recognition that while there is substantial research on cyberbullying detection in English, there is a noticeable gap in the literature regarding languages like Turkish. Given the unique linguistic characteristics of Turkish and the growing prevalence of social media use in Turkey, our study seeks to address this gap by developing a tailored model for the Turkish context.

Our motivation is driven by several key factors;

- Increasing Prevalence of cyberbullying; the rise of social media has unfortunately been accompanied by an increase in cyberbullying [24], which can have severe psychological impacts on individuals, especially young users.

- Linguistic Challenges; Turkish, with its rich morphological structure, presents unique challenges for natural language processing. Standard cyberbullying detection models, predominantly designed for English [25], are often ineffective for Turkish due to these linguistic complexities.

- Social Responsibility; as researchers, we feel a strong sense of responsibility to contribute to safer online environments. By developing a model that can accurately detect cyberbullying in Turkish, we aim to provide tools for social media, platforms and authorities to better protect users.

- Technical Innovation; we are motivated by the opportunity to advance the field of natural language processing and machine learning by tackling the explored area of Turkish language processing.

IV. PROPOSED FRAMEWORK AND APPROACH

In this scholarly inquiry, we advocate for an approach to architect a machine learning model tailored to identify instances of cyberbullying within Turkish tweets. A supervised learning approach has been utilized with two classifiers: Support Vector Machine (SVM) and Random Forest (RF). The methodology encompasses the following stages, as depicted in Figure 2:

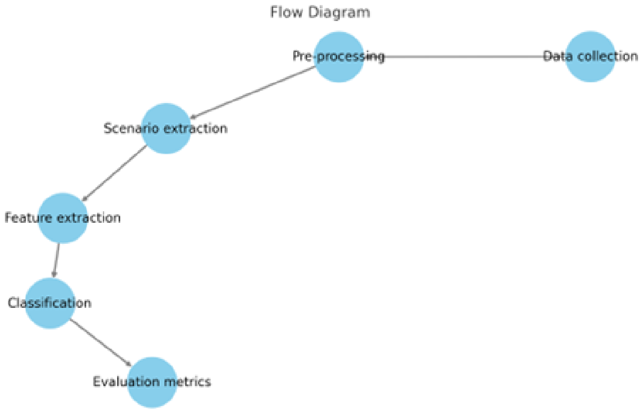

Figure 2. *Recommended methodology*

The problem definition of cyberbullying, including its global and Turkey prevalence and impact.

Global Prevalence and Impact of Cyberbullying

- Widespread Issue [26]; cyberbullying has emerged as a significant global concern with the increasing ubiquity of social media and online communication platforms.

- Psychological Effects [27]; it poses serious psychological risks, including anxiety, depression, and in extreme cases, suicidal thoughts, particularly among adolescents and young adults.

- Statistical Evidence [28]; studies have shown varying prevalence rates across different countries, with some reporting that up to 35-40% of young internet users have experienced cyberbullying.

Cyberbullying in Turkey

- Rising Concern [29]; in Turkey, the rise of internet and social media usage has been paralleled by an increase in cyberbullying incidents.

- Youth Vulnerability [30]; Turkish youth, who constitute a significant portion of the online population, are particularly susceptible to online harassment.

- National Studies and Statistics; recent surveys and studies in Turkey have indicated that cyberbullying is a growing problem, affecting a substantial number of adolescents. For instance, a study might reveal that over 30% of teenagers in Turkey have reported experiencing some form of online harassment [31].

- Cultural Context; the problem is compounded by cultural factors unique to Turkey, where social media is a vital part of daily life for many, yet discussions around cyberbullying and online safety are still emerging.

Problem Definition

- Broad Implications; cyberbullying encompasses various forms of online harassment, including spreading rumors, posting hurtful comments, and sharing private information without consent.

- Global and Local Relevance; while the problem is global, its manifestations can vary by culture and language, necessitating localized approaches to detection and prevention, as we propose in our study for the Turkish context.

Each of these stages is designed to collectively contribute to an effective and efficient system for detecting incidents of cyberbullying in Turkish tweets. We aim to navigate the nuances of the Turkish language using the Zemberek-NLP tool and to leverage machine learning techniques to accurately classify tweets. This approach is expected to enhance the effectiveness of cyberbullying detection and thus contribute to a broader goal of creating safer online environments.

A. DATA COLLECTION

The first step in our methodology involves data collection. On February 5, 2021, we gathered a dataset comprising 35,000 Turkish comments from Twitter APIs. These comments were subsequently labeled as 'bullying' or 'not bullying' based on the presence of bullying keywords that we manually collected from posts, which frequently occur in the Turkish community. This resulted in a comprehensive list of bullying words that we used to classify comments into bullying and non-bullying classes. Of the 35,000 comments collected [for example, 5,500 were labeled 'bullying', and the rest 'not bullying']. This distribution was taken into account when evaluating the model's performance. The collected comments include comments with both positive and negative emotional tones. Additionally, comments from users from different geographic regions, different age groups, and different genders are included in the dataset.

B. DATA PREPROCESSING - LABELING

A specific code set for Turkish is used for data preprocessing and tokenizing words. This allows a large text sample to be divided into individual words. In the data labeling process, a total of 3 experts were used.

The labeling process of the dataset was carried out using a two-stage approach:

- Automatic Keyword-Based Labeling: In the first stage, we created a list containing keywords related to bullying that frequently occurs in Turkish society. This list includes terms obtained from previous studies in the literature, as well as observations from social media and expert opinions. The collected 35,000 comments were automatically categorized into "bullying" and "not bullying" categories by filtering them through these keywords.

- Manual Review: After the automatic labeling process, a manual review was conducted on a randomly selected subset. This was done by 3 experts to validate and, if necessary, correct the accuracy of automatic labeling. During this process, comments that were misleadingly labeled were detected and corrected.

This two-stage approach was used with the aim of ensuring both speed and accuracy in labeling. While the automatic labeling process enables quick processing of large datasets, the manual review process helps enhance the accuracy of labeling.

In total, 35,000 Turkish comments were collected. Of these, 5,500 were labeled as 'bullying' and the remaining 29,500 as 'not bullying'. This labeling was based on a list containing signs of bullying and keywords. To better understand the effect of the preprocessing stage on the data, we examined a sample tweet in detail. The initial sample tweet was: "Today they treated me very badly at school. #badday". In the first step, we converted all characters to lowercase, resulting in "today they treated me very badly at school. #badday". Then, special characters, numbers, and emojis were removed from the tweet, giving us "today they treated me very badly in the school badday". In the third step, we eliminated frequently used stop words in Turkish (e.g., 'to me' and 'very'), obtaining "today school bad, treat bad". Finally, stemming was applied using Zemberek-NLP, and the final form became "today school bad, treat bad". These steps help make the text data moreprocessable for the model, aiding it in better understanding the text and thus making more accurate predictions. Prior to integrating the data into the model, it necessitates cleansing and normalization as delineated below:

B. 1. Data Cleaning

This process encompasses the exclusion of URLs, hashtags, "@" references, numerical values, non-Turkish lexemes, and extraneous components from the assembled tweets. Additionally, each word is refined through the following steps:

- Removal of repeated letters in words;
- Exclusion of prohibited terms, retaining words in textual form if they are not encompassed within the roster of disallowed words.

It is imperative to acknowledge that the Turkish linguistic framework is intricate, encompassing myriad grammatical constructs. Consequently, a singular term might manifest in diverse morphological variations, yet retain identical semantic values.

B. 2. Normalization

This process involves word normalization to eliminate potential letter misinterpretations. Due to common misspellings of some words, some Turkish letters have been replaced with their official forms.

B. 3. Zemberek-NLP

This toolkit offers an effective text processing solution for Turkish text. The tool provides a suite of capabilities, including tokenization, stemming, spell-checking, Named Entity Recognition (NER), Part-of-Speech tagging (POS tagging), and morphological analysis. Within the architectural design of our model, the stemming and tokenization functionalities were harnessed as delineated below:

B.3.1. Stemmed Words

This is the process of reducing inflected terms to their root forms. This is especially useful for Turkish, which has a rich morphological structure. This reduces the number of features and groups different forms of the same word together.

B.3.2. Tokenization

This entails a data delineation method that deconstructs a sentence into finer segments, commonly referred to as tokens. Each segment carries its own meaning. This enhances the granularity of our analysis and improves the perception of bullying terms.

B. 4. Extraction of Diverse Configurations

Subsequent to the preprocessing phase, we extrapolated six distinct scenarios from our dataset. These scenarios facilitate the exploration of various text processing strategies to optimize our model's performance:

B.4.1. Cleaned Data

This configuration encompasses data postprocessing, wherein URLs, hashtags, "@" references, numerical entities, non-Turkish lexemes, repetitive letters, and singular characters have been excised. Furthermore, this particular configuration has been subjected to normalization procedures.

B.4.2. Stemmed data

This scenario encompasses a cleaned dataset that underwent stemming using the Zemberek tool.

B.4.3. Tokenization

This scenario includes a cleaned dataset that underwent tokenization using the Zemberek tool.

B.4.4. Tokenization

This scenario includes a dataset cleaned by removing the encoded words.

B.4.5. Tokenization

This scenario includes data cleaned with the Zemberek tool and removed from the list of words to be blocked.

B.4.6. Tokenization

This scenario includes data cleaned with the Zemberek tool and removed from the list of words to be blocked. It is important to note that Turkish also has a rich long word sequence. Therefore, when these words are not considered, the performance of our model can usually be improved. After deriving these scenarios, both SVM and RF classifiers were used to determine which scenario provided the highest accuracy.

B. 5. Feature Extraction

At this juncture, the model metamorphoses the data into a structure conducive to the deployment of machine learning methodologies. The paramount aim is to extract salient attributes from the dataset. To realize this aim, we harnessed the capabilities of the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer in tandem with the Bag of Words (BoW) algorithm. These approaches allow us to extract the most essential features from the data and organize them in a feature list.

TF-IDF is a weight calculation approach commonly used in text mining. It assigns value to words collected through statistical analysis of a document's content. In this scholarly investigation, multiple TF-IDF analyzers, encompassing Unigram and Bigram, have been deployed.
Term Frequency (TF) typically undergoes normalization to ascertain its weight. Conceptually, it represents the occurrence frequency of a term (t) within a document (d), adjusted to the cumulative word count of the said document:

Formula (1) for Term Frequency (TF):

$$TF_{(t,d)} = \frac{n_{(t,d)}}{\sum t' \in d^{n(t',d)}} \tag{1}$$

Where:

- $TF_{(t,d)}$ is the term frequency of term *t* in document *d*.
- $n_{(t,d)}$ is the number of times term *t* appears in document *d*.
- $\sum_{t' \in d}{}^{n}(t',d)$ is the total number of terms in document *d*.

This formula correctly calculates the frequency of a term *t* in a document *d*, normalized by the total number of terms in that document. The removal of the index *k* clarifies the formula, ensuring it accurately represents the term frequency calculation as intended in the context of TF-IDF analysis.

Formula (2): Inverse Document Frequency (IDF)

The Inverse Document Frequency (IDF) is a measure of how much information a word provides, based on its frequency across all documents. It helps in assessing the significance of a word; less frequent words across documents are considered more significant. The formula for IDF is:

$$IDF(t) = \log(\frac{N}{df(t)} \tag{2}$$

Where:

- IDF(*t*) is the inverse document frequency of the term *t*.
- *N* is the total number of documents in the corpus.
- df(*t*) is the number of documents containing the term *t*.
- The logarithm scale is used to dampen the effect of IDF. It ensures that terms that appear in a small number of documents do not get an excessively high IDF value.

Formula (3): TF-IDF Weight

The TF-IDF weight is a statistical measure used to evaluate the importance of a word to a document in a collection or corpus. It is the product of two statistics, term frequency and inverse document frequency. The formula for calculating TF-IDF is:

$$TF - IDF(t, d) = TF_{(t,d)}\ x\ IDF(t) \tag{3}$$

Where:

- TF-IDF(t,d) is the TF-IDF score of the term tt in document d.
- TF$_{(t,d)}$ is the term frequency of term t in document d, as explained in Formula (1).
- IDF(t) is the inverse document frequency of the term t, as calculated in Formula (2).

The TF-IDF score increases proportionally to the number of times a word appears in the document, offset by the frequency of the word in the corpus. This helps to adjust for the fact that some words appear more frequently in general. TF-IDF is a key technique in text mining, used for text-based classifier algorithms like SVM and RF in our study.

The Bag of Words (BoW) serves as a representational technique that enumerates the recurrence of individual words within a text, yielding fixed-dimensional vectors. Each tweet is treated as a discrete data instance, and the occurrence rate of every term within that tweet is determined. This culminates in a representation termed as a vector, predicated upon the numerical quantification of the term. This method is effective in converting text data into a format that machine learning algorithms can process efficiently.

B. 6. Feature Extraction

After the feature extraction process, the collected dataset is randomly split into training and test sets in an 80:20 ratio. Concurrently, the training data serves as the foundation for instructing the model employing both SVM and RF classifiers, whereas the test dataset is leveraged to ascertain the model's efficacy in the concluding phase.

SVM (Support Vector Machine) is a supervised learning algorithm that can be used for both classification and regression tasks. It is especially effective in high-dimensional spaces and when the number of dimensions exceeds the number of samples. In addition to SVM, the Random Forest (RF) classifier has been used, which is an ensemble learning method that operates by constructing multiple decision trees during training and outputs the class, that is the mode of the classes of individual trees.

B. 7. Feature Extraction

Various classification metrics have been used to evaluate the analysis performance. The metrics of Accuracy (A), Recall (R), F1 score (F), and Precision (P) are deduced employing the subsequent mathematical expressions:

Precision (P):

$$P = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Positive\ (FP)} \tag{4}$$

Recall (R):

$$R = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Negative\ (FN)} \tag{5}$$

F1 Score (F):

$$F = 2.\frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Negative\ (FN)} \quad (6)$$

Accuracy (A):

$$A = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

True Positives (TP) represent the instances aptly identified as positive.

• True Negatives (TN) signify the instances accurately categorized as negative.
False Positives (FP) represent samples that are incorrectly classified as positive when they are actually negative.
• False Negatives (FN) denote the samples that are incorrectly classified as negative when they are actually positive.

V. ANALYSIS RESULTS AND DISCUSSION

We chose SVM due to its renowned effectiveness in text classification tasks, particularly in high-dimensional spaces like those encountered in natural language processing. SVM is adept at handling sparse data, which is common in textual analysis. Its ability to construct an optimal hyperplane for classification purposes makes it particularly suitable for differentiating between bullying and non-bullying content. Additionally, SVM's robustness in the face of overfitting, especially when dealing with limited training data, influenced our decision.

RF was selected for its proficiency in handling large datasets and its ability to maintain accuracy even when a significant portion of the data is missing or not clearly defined, which is often the case in social media content. RF's ensemble learning approach, where multiple decision trees are used to improve the predictive accuracy and control overfitting, makes it a powerful tool for classifying complex and noisy data such as tweets.

While other algorithms like Naïve Bayes, K-Nearest Neighbors, and Decision Tree are also used in similar contexts, we found SVM and RF to be more aligned with the specific challenges of our dataset, particularly considering the linguistic complexity of Turkish and the intricacies of cyberbullying detection.

Lastly, the selection of SVM and RF was based on their demonstrated strength in handling high-dimensional and noisy data, their robustness in classification tasks, and their suitability for the specific challenges presented by Turkish text data in cyberbullying detection.

Based on the results of our experiments, it has been determined that the SVM model combined with Zemberek-NLP shows superior performance in detecting Turkish cyberbullying tweets compared to other classifiers. Essentially, the performance in various experiments through various scenarios with the Zemberek-NLP tool via SVM and RF classifiers has been compared. Both of these classifiers are known for their effectiveness in text mining tasks. However, the analysis of the conducted experiments indicates that the SVM model combined with the stemming tool in the Zemberek-NLP toolkit yields the most promising results in the tested Turkish cyberbullying tweets. We undertook a comprehensive assessment of the outcomes from all experiments, synthesizing the findings and extrapolating insights from the paramount models across diverse scenarios. Based on the culled data, the scenario titled "Removal of blocked words from stemmed word data" manifested superior precision, registering an accuracy of 95.9% when paired with the SVM model facilitated by the TF-IDF vectorizer. Additionally, the SVM with BoW vectorizer showcased impressive performance with an accuracy of 95.7%. Interestingly, when distributed with NLTK, the "Removal of blocked words from stemmed word data" scenario provides the best accuracy. Within this framework, stemming serves to truncate inflected words to their foundational forms by excising suffixes, prefixes, and other affixations. This process augments the efficacy of pinpointing terms indicative of bullying. Hence, it can be concluded that the Zemberek toolkit improves accuracy and surpasses the most advanced data collection tools in Turkey. This is largely due to words having different structures, especially in Turkish. Finally, the model yields superior results when stop words are removed.

Table 1. *Synopsis of accuracy outcomes across varying test proportions.*

Scenario	0.2	0.4	0.6
Cleaned	94.2%	92.8%	91.8%
Stemmed words	94.9%	94.6%	93.6%
Discrete words	94.4%	94.9%	89.4%
Removing blocked words on cleaned data	94.2%	92.5%	89.4%
Removing blocked words on stemmed word data	93.7%	94.2%	89.1%
Removing blocked words on discrete word data	94.6%	93.1%	93.2%

Our findings show that the "Stemmed words" scenario with a test rate of 0.2 achieved 94.9% accuracy, closely followed by the "Removal of blocked words on discrete word data" scenario with 94.6% accuracy. However, when the test rate is increased to 0.6, the accuracy of the results generally decreases. Notably, a test ratio of 80:20 has given the most promising results. After settling on an 80:20 test ratio, we classified our dataset for further analysis using SVM and RF classifiers. Following this, stemming was undertaken using both the TF-IDF and BoW techniques. The BoW vectorizer emerges as an instrumental mechanism for extracting noteworthy features from textual repositories. It transfigures text-centric data into a matrix paradigm by gauging the frequency of terms within the designated corpus.

Table 2. *Compendium of peak accuracy outcomes for the delineated scenarios.*

Best Scenario Based on Accuracy {High to Low}	TF-IDF Ngram_Length (1, 2)	BoW Ngram_Length (1, 2)
Stemmed {With removed blocked words}	95.9%	95.7%
Stemmed	95.8%	95.2%
Discrete {With removed blocked words}	95.3%	95.0%
Discrete	94.8%	94.6%
Cleaned {With removed blocked words}	94.7%	94.5%
Cleaned	94.3%	94.1%

Table 3. *Confusion matrix for the SVM classifier*

Scenario	TN (%)	TP (%)	FN (%)	FP (%)
Cleaned	92.20	5.27	1.95	0.59
Stemmed	92.43	5.44	1.75	0.39
Discrete	92.26	5.30	1.90	0.54
Cleaned {With removed blocked words}	92.29	5.27	1.95	0.49
Stemmed {With removed blocked words}	92.39	5.57	1.70	0.34
Discrete {With removed blocked words}	92.30	5.36	1.85	0.49

Within the ambit of our empirical investigations utilizing the SVM classifier in conjunction with the TF- IDF vectorizer, the confusion matrix provides clarity on the distribution of True Negatives (TN), True Positives (TP), False Negatives (FN), and False Positives (FP) within our dataset. The "stem {Removal of blocked words}" scenario is determined to be the most effective in correctly classifying both non-bullying and bullying tweets. In the context depicted in Table 5, around 5.57% of tweets, originating from a pool of 537 bullying tweets, were aptly classified as cyberbullying. Concurrently, an estimated 92.39% of tweets were correctly ascertained from a total of 6524 non-bullying tweets. Therefore, the "stem {Removal of blocked words}" scenario is more effective in grading true negative and true positive results in the SVM model.

Table 4. *Efficacy Metrics for the Elucidated Scenarios*

Scenario	Class	Precision	Recall	F1 Score
Cleaned	No (0)	0.99050	0.988	0.989248
	Yes (1)	0.88770	0.862	0.874500
	Average	0.93910	0.925	0.931874
Stemmed	No (0)	0.99260	0.990	0.991298
	Yes (1)	0.91750	0.892	0.904600
	Average	0.95505	0.941	0.947949
Discrete	No (0)	0.99150	0.989	0.990248
	Yes (1)	0.90760	0.887	0.897200
	Average	0.94955	0.938	0.943724
Cleaned {With removed blocked words}	No (0)	0.99130	0.990	0.990650
	Yes (1)	0.89620	0.882	0.889100
	Average	0.94375	0.936	0.939875
Stemmed {With removed blocked words}	No (0)	0.99140	0.992	0.991700
	Yes (1)	0.90810	0.912	0.910300
	Average	0.94975	0.952	0.951000
Discrete {With removed blocked words}	No (0)	0.98990	0.990	0.989950
	Yes (1)	0.89120	0.890	0.890500
	Average	0.94055	0.940	0.940225

The conventional matrix of confusion was employed to deduce precision, recall, and F1 score across all delineated scenarios. As shown in Table 4, these measurements provide a comprehensive view of the performance of each scenario. Among the scenarios, the one using the Zemberek tool, "stemming data with removal of blocked words", achieves the highest accuracy. This particular configuration appears to be the most suitable for detecting Turkish cyberbullying. The Zemberek tool is especially effective due to its ability to reduce inflected terms to their root forms. Given that Turkish is a language with rich morphology, this presents a significant advantage. By removing prefixes, suffixes, and affixes, the detectability of definite bullying words has been enhanced, thus leading to an increase in the model's accuracy. Additionally, it's worth noting that the model's performance improves when stop words are removed. Stop words are common words that don't contribute to the meaning of a sentence and can be disregarded without loss of meaning. By removing these words, the model can focus on significant words that contribute to detecting cyberbullying.

Table 5. *Confusion matrix for the RF classifier.*

Scenario	TN (%)	TP (%)	TN (%)	TP (%)
Cleaned	76.53	4.37	1.62	0.49
Stemmed	76.72	4.52	1.45	0.32
Discrete	76.58	4.40	1.58	0.45
Cleaned {With removed blocked words}	76.60	4.37	1.62	0.41
Stemmed {With removed blocked words}	76.68	4.62	1.41	0.28
Discrete {With removed blocked words}	76.61	4.45	1.54	0.41

Our experiments resulting from the RF classifier using the TF-IDF vectorizer provide a breakdown of True Negatives (TN), True Positives (TP), False Negatives (FN), and False Positives (FP). The scenario "stem {Removal of blocked words}" has been determined to be particularly effective in correctly classifying both bullying and non-bullying examples of tweets. In this scenario, from Table 5, about 4.62% of tweets from 486 bullying tweets were correctly identified as cyberbullying, and approximately 76.68% of tweets from 5349 non-bullying tweets were correctly labeled. This observation underscores that the "stem {Removal of blocked words}" scenario exhibits particular adeptness in accurately categorizing both true negative and true positive outcomes within the RF model. Despite a decline in performance compared to the SVM model, the RF model maintains its relevance and utility in the field of cyberbullying detection.

Table 6. *Confusion matrix for the RF classifier in numerical terms.*

Scenario	TN	TP	FN	FP
Cleaned	5357	305	113	34
Stemmed	5370	316	101	22
Discrete	5360	308	110	31
Cleaned {With removed blocked words}	5361	305	113	28
Stemmed {With removed blocked words}	5367	323	98	19
Discrete {With removed blocked words}	5362	311	107	28

Table 6 presents the confusion matrix results obtained using the RF classifier for different preprocessing scenarios in numerical form. For each scenario, the True Negative (TN) values represent the number of examples where the model correctly classified comments that did not contain cyberbullying. The True Positive (TP) values indicate the number of examples where the classifier correctly identified comments containing cyberbullying. The False Negative (FN) values show the number of instances where the model misclassified comments that contain cyberbullying as non-cyberbullying, while the False Positive (FP) values indicate the number of times non-cyberbullying comments were misclassified as containing cyberbullying. These values are crucial for evaluating the model's performance and the impact of different preprocessing steps on the classifier.

Through our investigative lens, the SVM classifier has demonstrably surpassed the RF classifier in classification accuracy metrics. Within this scholarly exploration, we engaged in the classification of Turkish cyberbullying tweets leveraging these two machine learning methodologies. Significantly, the SVM classifier, harmoniously paired with the TF-IDF vectorizer, has demonstrated unparalleled proficiency in prognosticating cyberbullying comments. Comparatively, Mouheb et al. achieved an accuracy of 0.95 using the NB classifier. However, our methodology is improved compared to this result, and with the SVM classifier and the "Stemmed {Removal of blocked words}" scenario, an accuracy of 95.9% has been achieved. Table 7 delineates the accuracy metrics for both SVM and RF classifiers across diverse scenarios. The SVM classifier registers a pinnacle of accuracy at 95.9%, whereas the RF classifier attains an accuracy of 81.515% in the "Stemmed" context. Intriguingly, this contrasts sharply with the SVM classifier, wherein the "Stemmed {Removal of blocked words}" scenario proffers the most commendable outcomes. In conclusion, a hybrid classification approach might be the most effective solution when considering different contexts (scenarios).

Table 7. *Accuracy of classifiers for extracted scenarios.*

Scenario	SVM Accuracy (%)	RF Accuracy (%)
Stemmed {With removed blocked words}	95.9	81.430
Stemmed	95.8	81.515
Discrete {With removed blocked words}	95.3	81.005
Discrete	94.8	80.580
Cleaned {With removed blocked words}	94.7	80.495
Cleaned	94.3	80.155

VI. CONCLUSIONS

In this study, we trained the Support Vector Machine (SVM) model on a significant Turkish dataset consisting of approximately 35,000 comments. The model was later evaluated on a distinct Twitter dataset, selected in consideration of Twitter's pervasive utilization as a reservoir for textual data acquisition. Our primary aspiration was the adept classification of cyberbullying remarks. The results intimate that the efficacy of SVM, when allied with the TF-IDF vectorizer, is notably proficient in pinpointing instances of cyberbullying. These outcomes were contrasted with those procured from the Random Forest (RF) classifier, wherein we fine-tuned parameters such as the ngram range and incorporated auxiliary feature extraction techniques, notably the Bag of Words (BoW). It's important to note that BoW produces fixed-length vectors by counting the frequency of each word appearing in the text, using the CountVectorizer for this process. Despite these measures, SVM exhibited superior performance in detecting cyberbullying content, achieving an impressive accuracy of 95.9%. The high accuracy of our model offers a promising route to protect users from cyberbullying on social platforms.

Projecting ahead, we advocate for an expansive evaluation of our model, potentially spanning the analysis of millions of daily submissions across social media platforms. Furthermore, integrating this model within messaging applications could empower users with heightened cognizance of cyberbullying within social networks, simultaneously facilitating the automatic excision of comments laden with bullying. We can also envisage integrating a new version of the model with law enforcement and social aid organizations to monitor and address severe bullying incidents that lead to tragic outcomes like suicide.

A further research aim involves transitioning our model from traditional machine learning to deep learning techniques. This transition might provide an opportunity to compare the results produced by these two different training methodologies and could potentially provide more nuanced insights into cyberbullying detection.

VII. REFERENCES

[1] A. Mishrif and A. Khan, “Causal Analysis of Company Performance and Technology Mediation in Small and Medium Enterprises During COVID-19,” *Journal of the Knowledge Economy*, Oct. 2022,

[2] Erdal Özbay, “Transformatör-Tabanlı Evrişimli Sinir Ağı Modeli Kullanarak Twıtter Verisinde Saldırganlık Tespiti,” *Selcuk University Journal of Engineering, Science and Technology*, pp. 986–1001, Dec. 2022

[3] Ayça Balmumcu and Hilal Yüceyılmaz, “Investigation of Cyberbullying and Cyber Victimization Level of Young Women,” *Afyon Kocatepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi dergisi*, May 2023

[4] A. Blanchard and T. Horan, “Virtual Communities and Social Capital,” *Social Dimensions of Information Technology: Issues for the New Millennium*, 2000. https://www.igi-global.com/chapter/virtual-communities-social-capital/29107 (accessed Mar. 31, 2020).

[5] L. Cheng, K. Shu, S. Wu, Y. N. Silva, D. L. Hall, and H. Liu, “Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model,” *arXiv.org*, Aug. 06, 2020. https://arxiv.org/abs/2008.02642 (accessed Oct. 22, 2023).

[6] S. N. Firdaus, C. Ding, and A. Sadeghian, “Retweet Prediction based on Topic, Emotion and Personality,” *Online Social Networks and Media*, vol. 25, p. 100165, Sep. 2021

[7] S. C. S. Caravita, B. Colombo, S. Stefanelli, and R. Zigliani, “Emotional, psychophysiological and behavioral responses elicited by the exposition to cyberbullying situations: Two experimental studies,” *Psicología Educativa*, vol. 22, no. 1, pp. 49–59, Jun. 2016

[8] Rüstem Göktürk HAYLI and Yrd. Doç. Dr.yüksel ÇIRAK, “Siber Zorba Olan ve Olmayan Ergenlerin Yordanmasında Siber Mağduriyet, Akran Zorbalığı ve Karanlık Üçlünün Rolü,” *Journal of Inonu University Faculty of Education*, vol. 24, no. 1, pp. 420–448, May 2023

[9] K. Jordan, “From Social Networks to Publishing Platforms: A Review of the History and Scholarship of Academic Social Network Sites,” *Frontiers in Digital Humanities*, vol. 6, Mar. 2019,

[10] I. F. Kilincer, F. Ertam, and A. Sengur, “Machine Learning Methods for Cyber Security Intrusion Detection: Datasets and Comparative Study,” *Computer Networks*, vol. 188, p. 107840, Jan. 2021

[11] M. C. Martínez-Monteagudo, B. Delgado, Á. Díaz-Herrero, and J. M. García-Fernández, “Relationship between suicidal thinking, anxiety, depression and stress in university students who are victims of cyberbullying,” *Psychiatry Research*, vol. 286, p. 112856, Apr. 2020

[12] A. Muneer and S. M. Fati, “A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter,” *Future Internet*, vol. 12, no. 11, p. 187, Oct. 2020

[13] B. A. Talpur and D. O’Sullivan, “Cyberbullying severity detection: A machine learning approach,” *PLOS ONE*, vol. 15, no. 10, p. e0240924, Oct. 2020

[14] D. Olweus and S. P. Limber, “Some problems with cyberbullying research,” *Current Opinion in Psychology*, vol. 19, pp. 139–143, Feb. 2018

[15] H. Vandebosch and K. Van Cleemput, “Cyberbullying among youngsters: profiles of bullies and victims,” *New Media & Society*, vol. 11, no. 8, pp. 1349–1371, Nov. 2009

[16] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, “Social Media Cyberbullying Detection using Machine Learning,” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, 2019

[17] M. A. Al-Garadi *et al.*, “Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges,” *IEEE Access*, vol. 7, pp. 70701–70718, 2019

[18] A. Ali and A. M. Syed, “Cyberbullying Detection using Machine Learning,” *DOAJ (DOAJ: Directory of Open Access Journals)*, Sep. 2020

[19] E. Raisi and B. Huang, “Cyberbullying Detection with Weakly Supervised Machine Learning,” *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM ’17*, 2017

[20] M. Sadigzade and E. Nasibov, "Comparative Analysis of Count Vectorization vs TF-IDF Vectorization for Detecting Cyberbullying in Turkish Twitter Messages," in *Journal of Modern Technology & Engineering*, vol. 7, no. 1, 2022.

[21] B. ERDİ, E. A. ŞAHİN, M. S. TOYDEMİR, and T. DÖKEROĞLU, “Makine Öğrenmesi Algoritmaları ile Trol Hesapların Tespiti,” *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, Nov. 2020,

[22] V. Diogho and A. Paula, “Exploring Text Mining and Analytics for Applications in Public Security: an in-depth dive into a systematic literature review,” *Socioeconomic Analytics*, vol. 1, pp. 5–55, Jul. 2023

[23] A. Bozyigit, S. Utku, and E. Nasiboglu, “Cyberbullying Detection by Using Artificial Neural Network Models,” *2019 4th International Conference on Computer Science and Engineering (UBMK)*, Sep. 2019

[24] H. Baruah, P. Dashora, and M. K. Chaudhary, “Incidences of cyberbullying among adolescents,” *Advance Research Journal Of Socıal Scıence*, vol. 8, no. 2, pp. 143–149, Dec. 2017

[25] A. Al-Marghilani, “Artificial Intelligence-Enabled Cyberbullying-Free Online Social Networks in Smart Cities,” *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, Jan. 2022

[26] I. Aoyama and T. L. Talbert, “Cyberbullying Internationally Increasing,” pp. 183–201, Jan. 2010

[27] S. Skilbred-Fjeld, S. E. Reme, and S. Mossige, “Cyberbullying involvement and mental health problems among late adolescents,” *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, vol. 14, no. 1, Feb. 2020

[28] M.-J. Wang, K. Yogeeswaran, N. P. Andrews, D. R. Hawi, and C. G. Sibley, “How Common Is Cyberbullying Among Adults? Exploring Gender, Ethnic, and Age Differences in the Prevalence of Cyberbullying,” *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 11, pp. 736–741, Nov. 2019

[29] M. ERDOĞDU and M. KOÇYİĞİT, “The correlation between social media use and cyber victimization: A research on generation Z in Turkey,” *Connectist: Istanbul University Journal of Communication Sciences*, 2021

[30] Y. Akbulut and B. Eristi, “Cyberbullying and victimisation among Turkish University students,” *Australasian Journal of Educational Technology*, vol. 27, no. 7, 2011

[31] A. ARSLAN, O. BİLGİN, and M. INCE, “Lise Öğrencilerine Yönelik Siber Zorbalık Ölçeği Geliştirme Çalışması,” *OPUS Uluslararası Toplum Araştırmaları Dergisi*, pp. 1–1, Jun. 2020