Turkish Cyberbullying Detection with Fine-Tuned Pre-Trained Language Models

Araştırma Makalesi/Research Article

Metin BİLGİN[1*], Bilge Nur BEKAR[1]

[1]Bursa Uludağ Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye
metinbilgin@uludag.edu.tr , bilgenbekar@gmail.com
(Geliş/Received:05.08.2024; Kabul/Accepted:04.01.2025)
DOI: 10.17671/gazibtd.1528238

***Abstract*—** With the rapid increase in internet usage and its pervasive presence in all aspects of life, social media platforms have seen a rise in negative behaviors alongside their positive contributions. One such negative behavior is cyberbullying, which refers to the misuse of information and communication technologies to harm others. Cyberbullying is becoming a significant social problem. This study aims to detect and classify Turkish sentences containing cyberbullying using deep learning models. To achieve this, the BERT model, known for its ability to understand the context of language, was chosen. Specifically, the BERTurk, DistilBERTurk, and ConvBERTurk models—designed for the Turkish language—were fine-tuned and retrained using a dataset of 3,388 tweets labeled as racist, sexist, offensive language, or neutral. The primary goal of this study is to perform a comprehensive comparison of multi-class Turkish cyberbullying detection models and to develop an Artifical Intelligence (AI) model that delivers highly accurate results on real-world data. According to the results, BERTurk achieved the highest F1 score of 0.88, while the DistilBERTurk model showed the lowest performance.

***Keywords*—** natural language processing, transformers, BERT, cyberbullying, pre-trained language models

İnce Ayar Yapılmış Ön Eğitimli Dil Modelleri ile Türkçe Siber Zorbalık Tespiti

Özet— İnternet kullanımının hızla artması ve hayatın her alanında yaygın hale gelmesiyle birlikte, sosyal medya platformlarında olumlu katkıların yanı sıra bazı olumsuz davranışlar da artış göstermiştir. Bu olumsuz davranışlardan biri, başkalarına zarar vermek amacıyla bilgi ve iletişim teknolojilerinin kötüye kullanılmasını ifade eden siber zorbalıktır. Siber zorbalık, önemli bir toplumsal sorun haline gelmektedir. Bu çalışma, derin öğrenme modelleri kullanarak siber zorbalık içeren Türkçe cümleleri tespit etmeyi ve sınıflandırmayı amaçlamaktadır. Bu amaç doğrultusunda, dilin bağlamını anlama yeteneğiyle bilinen BERT modeli tercih edilmiştir. Özellikle, Türkçe dilini destekleyen BERTurk, DistilBERTurk ve ConvBERTurk modelleri, ırkçı, cinsiyetçi, saldırgan dil veya nötr olarak etiketlenen 3.388 tweet içeren bir veri kümesiyle ince ayar yapılarak yeniden eğitilmiştir. Çalışmanın temel hedefi, çok sınıflı Türk siber zorbalığını tespit eden modellerin kapsamlı bir karşılaştırmasını yapmak ve gerçek dünya verileri üzerinde yüksek doğrulukla sonuçlar üreten bir yapay zeka modeli geliştirmektir. Sonuçlara göre, BERTurk 0,88 F1 puanı ile en yüksek başarıyı elde ederken, DistilBERTurk modeli en düşük performansı göstermiştir.

***Anahtar Kelimeler*—** doğal dil işleme, transformers, BERT, siber zorbalık, ön eğitimli dil modelleri

1. INTRODUCTION

The internet continues to permeate every aspect of our daily lives. Its widespread use directly impacts social life as well. Information technologies, with their advantages, have led to various changes in our way of life [1]. For instance, many areas such as shopping, communication, education, and entertainment are now actively conducted online. One of the areas where the internet is used most in recent times is social media platforms. Through social media, individuals can easily reach large audiences, assume desired identities, and engage in various activities. While there are advantages to this, the intensive use of the internet has also led to the proliferation of behaviors such as joy, sadness, anger, and bullying within social communication networks [2]. The misuse of information and communication technologies to harm others is referred to as cyberbullying [3]. The concept was first coined by Bill Belsey [3]. Cyberbullying refers to the reflection of harmful behaviors such as insults, ridicule, racist remarks, and judgmental attitudes on social networks. The rate of cyberbullying is increasing rapidly worldwide, and according to 2021 data, it has a rate of approximately 16% among types of bullying. [4]. The anonymity of online communication fosters a belief that actions will go unpunished, contributing to the growing prevalence of this phenomenon [5]. The platforms where cyberbullying is most frequently observed include social networking sites such as Facebook and Twitter [6]. The rapidly increasing presence of cyberbullying on the internet poses threats to societies. Therefore, detecting cyberbullying is crucial to mitigating its harmful effects. The AI-based cyberbullying detection project aims to utilize AI models to provide reliable detection of cyberbullying.

Numerous studies have been conducted on cyberbullying in recent years, with many employing traditional machine learning methods [7][8][9]. These studies have explored various aspects of cyberbullying detection, yet they often fall short in terms of addressing the complexities introduced by nuanced language and the diversity of cyberbullying behaviors. However, the growing use of deep learning models in natural language processing (NLP) can be attributed to advances in hardware, the increasing volume of available data, and the rise of open-source projects. Recent innovations in deep learning, particularly the development of transformative architectures, have significantly enhanced NLP by improving both the speed and accuracy of word context understanding. Despite these advancements, a universally effective tool for detecting cyberbullying has yet to be developed. As a result, further research in this area is essential.

A review of the existing literature reveals a lack of Turkish-specific data related to cyberbullying, with most available datasets focusing on binary classification. Our study seeks to fill this gap by applying multi-class classification techniques using widely-used, high-performance pre-trained models such as BERT, DistilBERT, and ConvBERT. We evaluate the performance of these models by calculating precision, recall, confusion matrix, PR curves, and F1 score metrics. Additionally, user interaction is incorporated into this process. The system we developed generates appropriate labels and scores to indicate whether the input text contains instances of cyberbullying.

The objectives of this study are as follows:

- To achieve a comprehensive study by performing multi-class classification for cyberbullying detection using a multi-class dataset.
- To fine-tune and train deep learning-based pre-trained models and compare the performance of the developed models based on evaluation metrics.
- To contribute to research in Turkish natural language processing, particularly in text classification and cyberbullying detection. Turkish-supported BERT models, which have been insufficiently explored in the literature, using a multi-class dataset.
- To enhance cyberbullying awareness through the evaluation of data containing cyberbullying.
- To develop a system that provides users with an easy way to detect Turkish cyberbullying.

2. LITERATURE REVIEW

The study focuses on a multi-class text classification problem, with an emphasis on Turkish research and recent developments in the field. The methods discussed in the literature review are categorized under relevant headings and critically analyzed in the context of this study. There are many machine learning methods in the literature. In addition, studies involving deep learning methods are also included in the literature review. In contrast, our research contributes to the existing literature by using transformer-based architectures, known for their high performance in natural language processing tasks.

In this study, the methods are presented by incorporating changes based on the accumulated knowledge obtained from the studies observed in the literature. A dataset containing neutral, offensive language, sexism, and racism was selected to comprehensively address the detection of cyberbullying in Turkish. A multi-class Turkish dataset was used to implement BERT-based models, and BERTurk, ConvBERTurk, and DistilBERTurk models were applied to develop a Turkish cyberbullying detection system. In this study, for the first time in the literature, BERTurk, ConvBERTurk, and DistilBERTurk models were evaluated using the Turkish multi-class cyberbullying dataset, using the F1 score, recall, precision, confusion matrix, and PR curve.

2.1 Machine Learning Approaches in Existing Literature

Sevli and Sezgin used machine learning methods to detect and categorize cyberbullying in social media posts. The study used a dataset consisting of 47,692 English tweets. The dataset was balanced and contained six classes: bullying based on belief, gender, age, ethnicity, other characteristics, and non-cyberbullying. Confusion matrix,

F1 score, precision, and recall measures were used to compare KNN, SVM, and Random Forest algorithms. The best result was obtained with the SVM model, which achieved 83% accuracy. The best-performing category was non-cyberbullying tweets. It was suggested that the relatively smaller size of this class (16%) and the presence of fewer meaningful expressions could explain this result. The study suggested the use of deep learning techniques to address the problem more effectively [7]. This study was examined because it classified cyberbullying in multi-class, and the dataset contained similar categories. Although the balanced dataset provides a high accuracy rate with machine learning methods, the fact that the dataset is in English limits its applicability in detecting multi-class cyberbullying in Turkish, which has not yet been widely addressed in the literature.

In their study, Rohini and Ramchander focused on detecting cyberbullying in digital forums using machine learning methods. Two datasets were employed. The first dataset contains 10,000 comments, with 80% of the data not involving cyberbullying and 20% involving it. The second dataset includes 20,000 comments, where 60% of the data does not contain cyberbullying and 40% does, making it an imbalanced dataset. The best results were achieved using the Random Forest approach, with an accuracy of 99% [8]. Despite the imbalance, high performance was observed across both datasets. The study demonstrates high effectiveness in detecting cyberbullying, even with the imbalanced nature of the dataset.

Bozyiğit et al.'s study emphasizes the role of social media features in cyberbullying detection, creating a balanced dataset of 5,000 labeled posts and using the chi-square test to explore the relationship between features like the sender's follower count and cyberbullying. They tested machine learning algorithms on two datasets: one with only text features and the other with social media features. The latter showed better performance [9]. This binary classification study contributes to Turkish cyberbullying detection but differs methodologically from ours, as we use BERT-based deep learning models, multi-class classification, and a broader set of performance metrics. Both studies advance the field, yet our research highlights the evolution from traditional machine learning to deep learning and provides deeper insights into different types of cyberbullying.

In this study, Çöltekin developed a dataset to automatically detect offensive language in Turkish social media posts and conducted experiments using various machine learning methods. His primary focus was the classification of offensive language. The dataset was derived from Turkish tweets posted between 2018 and 2019, consisting of 36,232 manually labeled tweets. Each tweet was classified as either offensive or non-offensive, with a third annotator making the final decision in cases of disagreement. Çöltekin also evaluated the dataset in terms of cyberbullying analysis across different regions of Turkey and assessed the general rate of tweets containing cyberbullying, offering valuable insights. He employed SVM, n-gram, and BM25 models to automatically classify offensive tweets. The study achieved an F1 score of 77.3% for identifying offensive tweets, 77.9% for determining whether a specific tweet was targeted, and 53.0% for classifying targeted offensive tweets into three subcategories [10]. The study also addressed the multi-label classification problem. In contrast, our study tackles the multi-class classification problem with greater success.

2.2 Hybrid and Advanced Approaches

Sel and Hanbay explored various algorithms, including machine learning-based TFIDF+SVM, deep learning-based CNN and LSTM, and pre-trained Turkish language models like BERT, DistilBERT, and Electra. Their study focused on binary text classification using a partially balanced dataset of 5,292 tweets, evaluating models based on accuracy, F1 score, specificity, and sensitivity. BERT achieved the highest accuracy at 80%. In gender determination, the SVM classifier was compared to pre-trained language models. Although SVM requires more parameters and doesn't account for word meanings, it performed similarly to pre-trained models when contextual understanding was less important. While the study is valuable for exploring different models, further research is needed to achieve higher accuracy with balanced and larger datasets [11].

Nergiz and Avaroğlu trained an LSTM neural network with three different word embedding models using a dataset of 180,000 comments collected from three different social media platforms and set the epoch value to 10. The model using the Fasttext method achieved the highest accuracy of 93%. It was evaluated that the factors contributing to this high accuracy were the use of balanced data for binary classification and the implementation of data preprocessing steps. The limited increase in success was attributed to the non-standardized structure of social media spelling rules and the insufficient content of the comments coming from the Instagram platform [12]. The study stands out in terms of evaluating the adequacy of social media data. The dataset we used in the study shed light on the interpretation of the adequacy of social media data.

2.3 BERT and Transformer Model Approaches

Karaman used the BERT model for the binary classification of discriminatory-exclusionary tweets against Syrian refugees in his study. The study utilized the pre-trained BERT-BASE-TURKISH-UNCASED model, trained on Turkish documents. The dataset consisted of 2,264 tweets, and the model was trained for 12 epochs, achieving an accuracy of 0.8562 during training and 0.81 on the test set. It was suggested that increasing the sample size could enhance the model's sensitivity and accuracy [13]. While the study is similar to ours in using the BERT model for cyberbullying detection, there are key differences, such as the use of different models and a focus on binary classification.

In this study, Beyhan et al. conducted experiments on both binary and multi-class classification problems using the Istanbul Convention and Refugees datasets. They retrained the model with a 5-class dataset representing hate speech, aiming to accurately classify data on topics like hate speech and cyberbullying using various text classification techniques. The BERTurk model was employed, and results were evaluated using 5-fold cross-validation. On the Istanbul Convention dataset, the binary classification achieved an average accuracy of 77.06%, and multi-class classification reached 72.22%. The F1 scores were 77.86% and 72.22%, respectively. However, challenges arose as the BERT model, trained on formal sources, struggled with informal and short texts like those from Twitter. Additionally, the unbalanced nature of the Istanbul Convention dataset impacted classification results [14]. This study offers a comprehensive approach to detecting both multi-class and binary Turkish cyberbullying. In contrast, our research addresses the gap by comparing the performance of different BERT-based models, achieving a higher F1 score using a different dataset. Furthermore, while this study focused on tweets related to specific events (such as the Istanbul Convention or refugee issues), our study was trained on tweets from a dynamic, independent cyber environment.

In their study, Çelikten and Bulut developed a model using the BERT model to classify ten diseases in a dataset consisting of Turkish medical texts. BERTurk models were preferred due to the fact that Turkish is an agglutinative language and the morphological difficulties it presents in natural language processing. In addition, a multilingual BERT model developed by Google was used. The evaluation metrics of the study included precision, recall, F1 score, and weighted and macro averages. It was seen that the BERTurk model outperformed the multilingual BERT model [15]. This study shows that the BERTurk model stands out in terms of comparing BERT models supporting Turkish. The study provides resources for examining the BERTurk model with evaluation metrics and examining its performance. Aytan and Sakar conduct a comparative analysis of transformer-based models for Turkish natural language processing problems in their study. The study examined BERT, ConvBERT, and Electra models on sentiment analysis, named entity recognition (NER), and text classification problems using pre-trained Turkish models, while the RoBERTa model was trained with a large Turkish corpus. For text classification with a seven-category dataset, the BERTurk model achieved the highest classification performance with an accuracy of 94%. The ConvBERT model achieved a performance of 93.9% [16]. The study provides us with resources to examine the use of BERTurk and ConvBERTurk models in text classification problems.

In their study, Özkan and Görkem apply the BERT model to a multi-class classification problem. They preprocess the dataset with steps such as tokenization, case transformation, removal of stop words, deletion of numbers, and stemming, and then apply two different training-validation ratios to the dataset. The first ratio is 80% training and 20% testing, while the second is 85% training and 15% testing. The AdamW optimization method was selected as the optimizer. It was observed that AdamW showed less overfitting compared to models trained with the Adam optimization algorithm. The model with an 85% training ratio achieved an F1 score of 96%. The 85% training ratio gave better results compared to the 80% training ratio [17]. The study guides the use of AdamW optimization and the BERT model in a multi-class classification problem.

Arzu and Aydoğan performed a comparative analysis of BERTurk models for Turkish sentiment classification. In their study, they achieved the highest accuracy of 83% using the preprocessed ConvBERTurk mc4 (without case) model with a balanced dataset of 150,000 samples. The lowest performance was observed in the DistilBERTurk cased model [18]. The models used in their study were also applied to cyberbullying detection in our research. The study provided insight into the BERTurk, ConvBERTurk, and DistilBERTurk models.

Table 1. Literatıre Review Summary

Study	**Methods/Approaches**	**Classification Type**	**Dataset Language**	**Topic**	**Best Performance**
Sevli & Sezgin [7]	SVM, KNN, Random Forest	Multi-class	English	Cyberbullying detection (6 categories, balanced)	SVM: 83% accuracy
Rohini & Ramchander [8]	Random Forest	Binary	English	Cyberbullying detection (imbalanced 2 datasets)	Random Forest: 99% Accuracy
Bozyiğit et al. [9]	Traditional ML	Binary	Turkish	Cyberbullying detection with social media features	ML with social features: High correlation
Çöltekin [10]	SVM, BM25	Binary	Turkish	Offensive language detection (multi-label)	SVM: 77.9% F1 score (targeted tweets)
Sel & Hanbay [11]	TFIDF+SVM, CNN, LSTM, BERT, DistilBERT, Electra	Binary	Turkish	Text (gender) classification with pre-trained models	BERT: 80% accuracy

Nergiz & Avaroğlu [12]	LSTM with Fasttext	Binary	Turkish	Classification of Cyberbullying in Social Media Comments	Fasttext+LSTM: 93% accuracy
Karaman [13]	BERTurk (BASE-TURKISH-UNCASED)	Binary	Turkish	Discriminatory tweets about Syrian refugees	BERTurk: 81% test accuracy
Beyhan et al. [14]	BERTurk	Multi-class/Binary	Turkish	Hate speech and cyberbullying detection	BERTurk: 77.06% accuracy (binary)
Çelikten & Bulut [15]	BERTurk, Multilingual BERT	Multi-class	Turkish	Medical text classification (10 diseases)	BERTurk: Outperformed Multilingual BERT
Aytan & Sakar [16]	BERTurk, ConvBERT, Electra	Multi-class	Turkish	Sentiment analysis and text classification	BERTurk: 94% accuracy
Özkan & Görkem [17]	BERT	Multi-class	Turkish	Multi-class classification with optimization	BERT+AdamW: 96% F1 score
Arzu & Aydoğan [18]	ConvBERTurk mc4, DistilBERTurk	Binary	Turkish	Sentiment classification with balanced data	ConvBERTurk mc4: 83% accuracy

Upon examining the existing literature, it is evident that various approaches have been employed in the field of cyberbullying detection. An evolution from machine learning methods to deep learning techniques, particularly transformer-based models, can be observed. However, significant gaps still exist in the area of Turkish cyberbullying detection.

Firstly, studies on multi-class cyberbullying detection in the Turkish language are limited. Most existing research has focused on binary classification problems. In this context, our study aims to fill this gap in the literature by adopting a multi-class approach (neutral, offensive language, sexism, and racism).

Secondly, there is a lack of comparative analysis of the performance of BERT models specifically developed for the Turkish language in cyberbullying detection. Our study evaluates the effectiveness of BERTurk, ConvBERTurk, and DistilBERTurk models in Turkish cyberbullying detection by comparing them on the same dataset. This comparison allows us to determine which model is most suitable for this task, providing a valuable reference point for future research.

Thirdly, most existing studies have evaluated the performance of the models used with limited metrics. Our study employs a comprehensive set of metrics, including F1 score, recall, precision, confusion matrix, and PR curve, to evaluate the models' performance from multiple angles. This approach enables a more in-depth understanding of the strengths and weaknesses of the models.

Finally, the methodological innovation of our study lies in its systematic comparison of different BERT-based models' performance in Turkish cyberbullying detection. This comparison not only identifies the best-performing model but also reveals the success of each model in detecting different types of cyberbullying. This detailed analysis can guide future studies and assist researchers in selecting the most appropriate model for cyberbullying detection in Turkish natural language processing.

In conclusion, our study aims to fill existing gaps in the field of Turkish cyberbullying detection, comparatively evaluate the performance of the most up-to-date BERT-based models, and improve methodological approaches in this area. The findings of this study will contribute to the advancement of research in the field of Turkish cyberbullying detection by providing a solid foundation for future investigations.

3. MATERIAL AND METHODS

This section will provide information about the materials and methods used in the study.

The stages of the study are outlined below:

- Three distinct language models were trained on the same dataset to classify sentences containing racism, sexism, offensive language, and neutral expressions related to cyberbullying.
- The trained models and tokenizer files were shared on the HuggingFace platform, providing resources and opportunities for users to test the model for Turkish cyberbullying detection.
- To ensure rapid and reliable interaction of the cyberbullying detection with software, an API was developed using FastAPI.
- A user-friendly and lightweight interface was developed to facilitate easy cyberbullying detection.
- This study contributes to the field of Turkish natural language processing, which has limited research, by providing reliable and high-accuracy results in cyberbullying detection using deep learning-based models.

The flow diagram of the study can be seen in Figure 1.

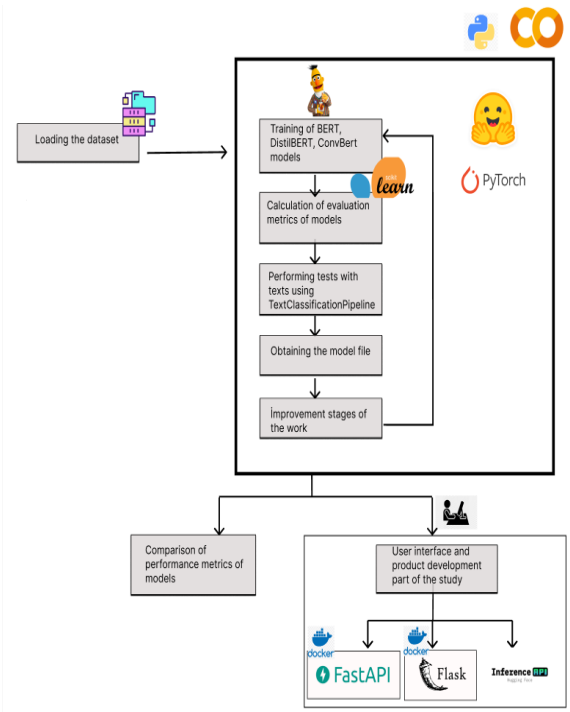

Figure 1. Flow diagram of the study

3.1. Dataset

Finding a labeled, multi-class dataset for Turkish cyberbullying is quite challenging. In this study, the Turkish-social-media-offensive-bullying dataset [19] provided by the HuggingFace platform was chosen, as it is the only available dataset. This dataset consists of 3,388 pre-processed samples, labeled for cyberbullying detection across four classes: racism, sexism, offensive language, and neutral. These classes cover significant aspects of cyberbullying detection and represent the only multi-class labeled dataset available in this domain. To address the imbalance in the dataset, data reduction techniques were applied to the most populated class, Neutral. The dataset contains 490 samples for Racism, 601 for Sexism, 910 for Offensive Language, and 980 for Neutral.

The initial and final distribution graphs of the dataset are shown in Figure 2.

a. b.

Figure 2. Distribution Graph of Classes

a. Distribution graph of the old dataset b. Distribution graph of the new dataset

Word clouds of the classes of the data set are shown in Figure 3.

a. b.

c. d.

Figure 3. Word clouds for each class: a. Word cloud for the Sexism class b. Word cloud for the Racism class c. Word cloud for the Offensive language d. Word cloud for the Neutral class.

One of the main reasons for choosing this dataset is that it contains four key classes that help detect cyberbullying. It is also the only Turkish open-source dataset with these categories. The Offensive Turkish Dataset [20], another open-source, multi-class cyberbullying dataset, provides an ideal resource for studies focused on multiple classifications and labeling. Cyberbullying data can be categorized into several types: non-aggressive language with swearing or untargeted attacks, attacks against a group, individual attacks on a person, and attacks targeting a non-human entity, such as an event or organization. Additionally, data can belong to more than one class. For these reasons, we decided to proceed with the Turkish-social-media-offensive-bullying dataset for our study.

3.2. Pre-Trained Models

This section provides information about the models used in the study.

1) BERT: Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model based on the evolving transformer architecture. The BERT model generates numerical representations by evaluating context-appropriate words. Specifically, the numerical vectors of words with similar meanings are closely aligned, whereas vectors for different meanings of the same word are less similar. BERT's architecture simultaneously considers both left and right contexts, which contributes to its effectiveness in interpreting word meanings even in complex language processing tasks [21]. The BERT model is composed of a 12-layer transformer structure [22]. The architecture of the BERT model is presented in Figure 4.

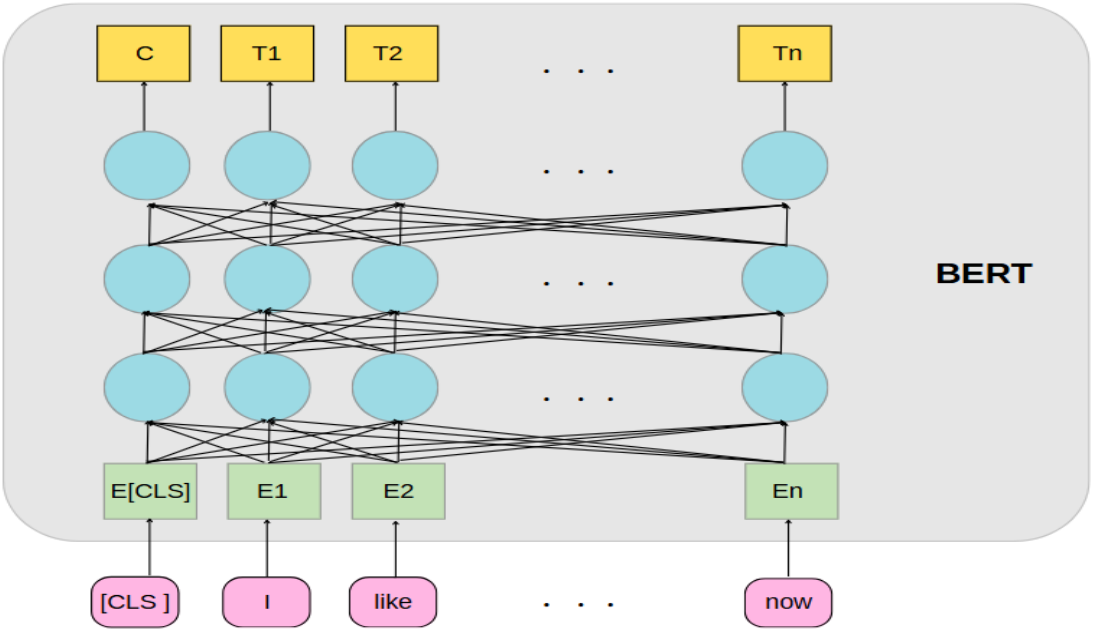

Figure 4. BERT model structure [21]

2) ConvBERT: The ConvBERT model is developed by replacing the self-attention layer of the BERT architecture with a span-based dynamic convolution. While the self-attention layer in the BERT model requires querying all inputs, the span-based self-attention allows for the examination of local dependencies. This modification aims to reduce the high memory usage and computational cost associated with the self-attention layer in BERT, thereby achieving better performance [16]. The ConvBERT model maintains the general structure of BERT while enhancing the attention mechanism by introducing convolution, thus improving the model.

ConvBert avoids the bottleneck of the BERT model and uses mixed attention. It is important to emphasize that ConvBERT uses both self-attention and convolution mechanisms. This shows that combining the two approaches yields better results [23]. The approaches presented in Figure 5 are discussed.

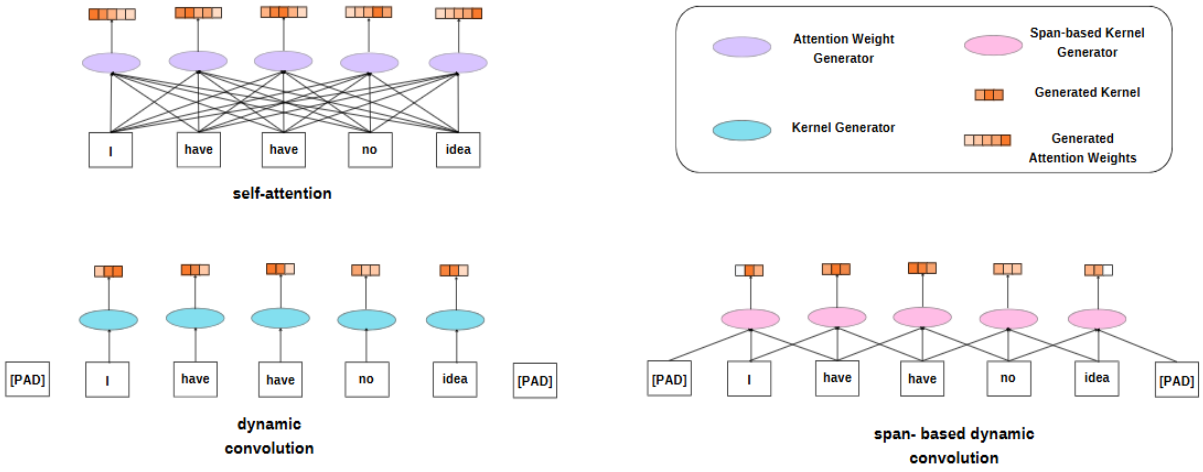

Figure 5. Approaches of self-attention, dynamic convolution, and span-based dynamic convolution [24]

3) DistilBERT: DistilBERT, introduced in a 2019 paper [24] and developed by Hugging Face [25], is a distilled version of larger models like BERT, designed to address the challenge of limited computational resources. The model claims a 40% reduction in size while retaining 97% of the linguistic understanding of BERT and increasing processing speed by 60% [24]. Its compact nature makes it especially well-suited for deployment on mobile devices and real-time applications. DistilBERT has fewer transformer layers compared to BERT and, in some studies, has been observed to have lower accuracy than BERT [26].

4) BERTurk: The pre-trained transformer models used in this study were pre-trained on Turkish data using a specialized corpus provided by Kemal Oflazer, a filtered version of the Turkish OSCAR corpus Wikipedia dumps, and various OPUS corpora [27]. The models employed in the study include bert-base-turkish-uncased, distilbert-base-turkish-cased, and convbert-base-turkish-mc4-cased, all trained with the specified datasets.

3.3. Fine-Tuning Training Phase

These models were trained for Turkish cyberbullying detection in a Google Colab environment, using a Tesla T4 GPU with CUDA. The GPUs enabled parallel computation, which allowed the models to be fine-tuned more quickly and with better memory management. The Turkish-social-media-offensive-bullying dataset was split into 80% for training and 20% for testing. To address the dataset imbalance, where the "Neutral" class was the most frequent with 1387 samples, the number of samples was reduced to 980 for better results. The distribution of word counts, including the minimum, maximum, and average counts, was also examined through a box plot (Figure 4). Based on this analysis, a maximum length (max_len) of 100 was selected to accommodate subword tokenization in the models.

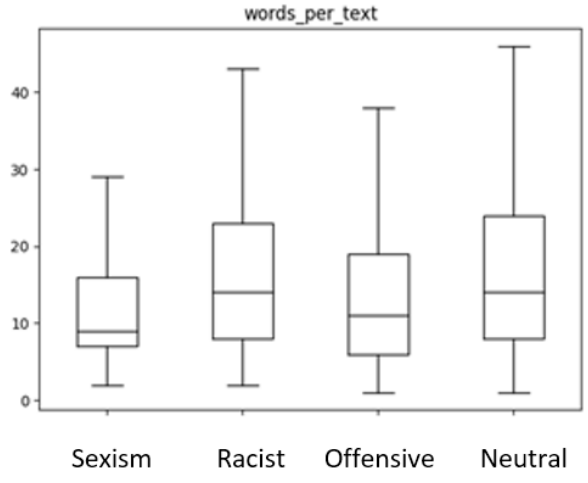

Figure 6. Box plot of word counts by class

The BERTurk model was trained using the pre-trained bert-base-turkish-uncased model and tokenizer through the Transformers library with 7 epochs and a batch size of 16. The AdamW algorithm was employed for optimization. The created model was trained using the initial weights of the BERTurk model. For training the DistilBERTurk model, 10 epochs and a batch size of 16 were used. The ConvBERTurk model was trained with 9 epochs and a batch size of 16. Although the training duration of the ConvBERTurk model is similar to that of the BERTurk model, it is somewhat shorter.

3.4. Evaluation Metrics

Precision: Precision, also known as sensitivity, measures how many of the positive predictions are correctly identified as true positives.

$$\text{Precision} = TP/(TP+ FP) \quad (1)$$

Recall: Recall, also known as sensitivity, measures the proportion of actual positive instances that are correctly identified as true positives in classification.

$$\text{Recall} = TP/(TP+FN) \quad (2)$$

F1 Score: The F1 score is used to measure classification performance. It is computed as the harmonic mean of precision and recall values. The F1 score is particularly useful in cases where the distribution is imbalanced.

$$F1 = 2 \times ((PxR)/(P+R)) \quad (3)$$

Confusion Matrix: The confusion matrix provides a summary of a classification problem by comparing the true labels with the predicted labels [26].

Precision-Recall Curves (PR Curves): It allows for the examination of the performance of precision and recall values in models with class imbalance

4. APPLICATION AND RESULTS

In the study developed for the detection of cyberbullying in Turkish, the dataset was divided into training and test data in equal proportions (80%-20%). The results obtained from the BERTurk, DistilBERTurk, and ConvBERTurk models were compared. The confusion matrices regarding the class prediction performance in the test dataset are presented in Figures 7-9. When the confusion matrices were examined, it was revealed that the models showed more errors in predicting the Offensive Language and Neutral classes. The ConvBERTurk and BERTurk models produce similar results. The results of the models are given in detail in Tables 2-4. The study achieved high success in the detection of cyberbullying in Turkish with the multi-class transformative model by obtaining an F1 score of 0.88. When the research results and project development process were evaluated, it was seen that reducing the Neutral category reduced the bias in the model outputs and improved the generalization ability. 2385 training samples and 596 test samples were used for model training. Among

the BERTurk, DistilBERTurk, and ConvBERTurk models, the BERTurk model achieved the highest F1 score of 0.884, while the DistilBERTurk model achieved the lowest F1 score of 0.83. PR curves are given in Figure 10 to examine the effects of class imbalance on the model and its performance. The graphs show that the Sexism class achieved the highest performance in all models with balanced precision and recall values. High locality and steep curves show good performance in accurate predictions and identifying positive examples. The BERTurk model is the most balanced in terms of class distribution, followed by the ConvBERTurk model. Comparing the balanced structure of the PR curve is useful in comparing the success of the models.

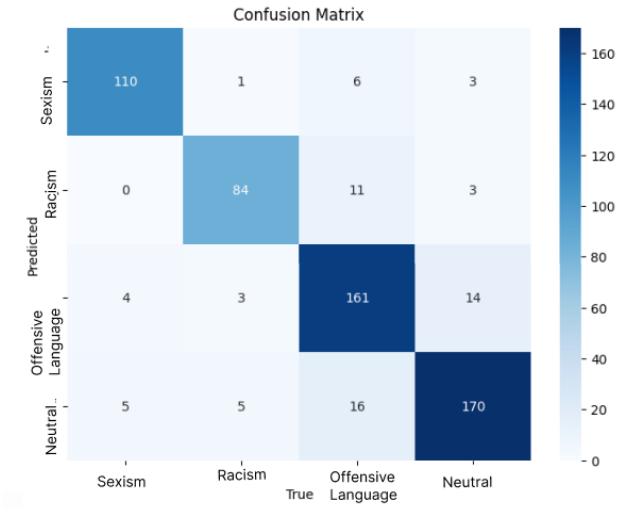

Figure 7. BERTurk confusion matrix

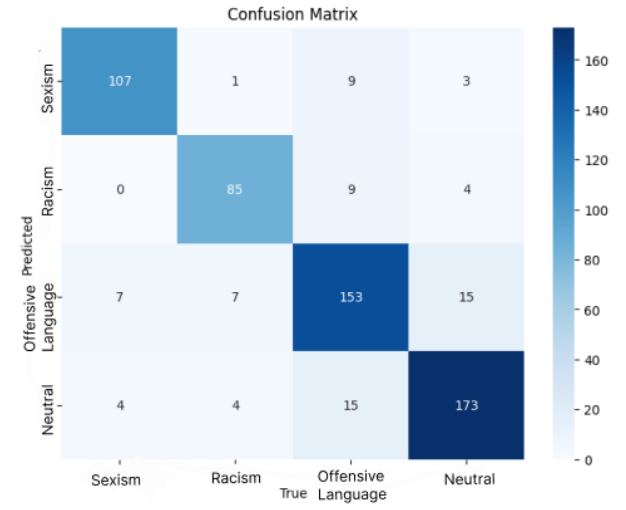

Figure 8. ConvBERTurk confusion matrix

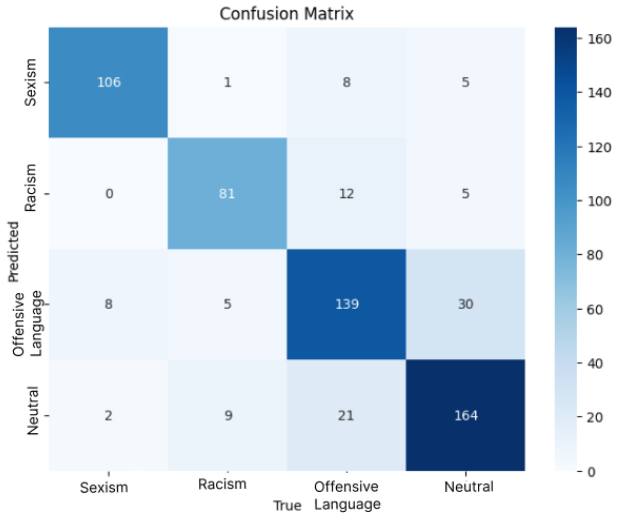

Figure 9. DistilBERTurk confusion matrix

Table 2. BERTurk model evaluation results

Classes	Precision	Recall	F1 score
All Classes	0.888	0.881	**0.884**
Sexism	0.924	0.917	**0.920**
Racist	0.903	0.857	0.880
Offensive	0.830	0.885	0.856
Neutral	0.895	0.867	0.881

Table 3. ConvBERTurk model evaluation results

Classes	Precision	Recall	F1 score
All Classes	0.873	0.871	0.872
Sexism	0.907	0.892	**0.899**
Racist	0.876	0.867	0.872
Offensive	0.823	0.841	0.831
Neutral	0.887	0.883	0.885

Table 4. DistilBERTurk model evaluation results

Classes	Precision	Recall	F1 score
All Classes	0.833	0.828	0.830
Sexism	0.914	0.833	**0.898**
Racist	0.844	0.826	0.835
Offensive	0.772	0.764	0.768
Neutral	0.804	0.837	0.820

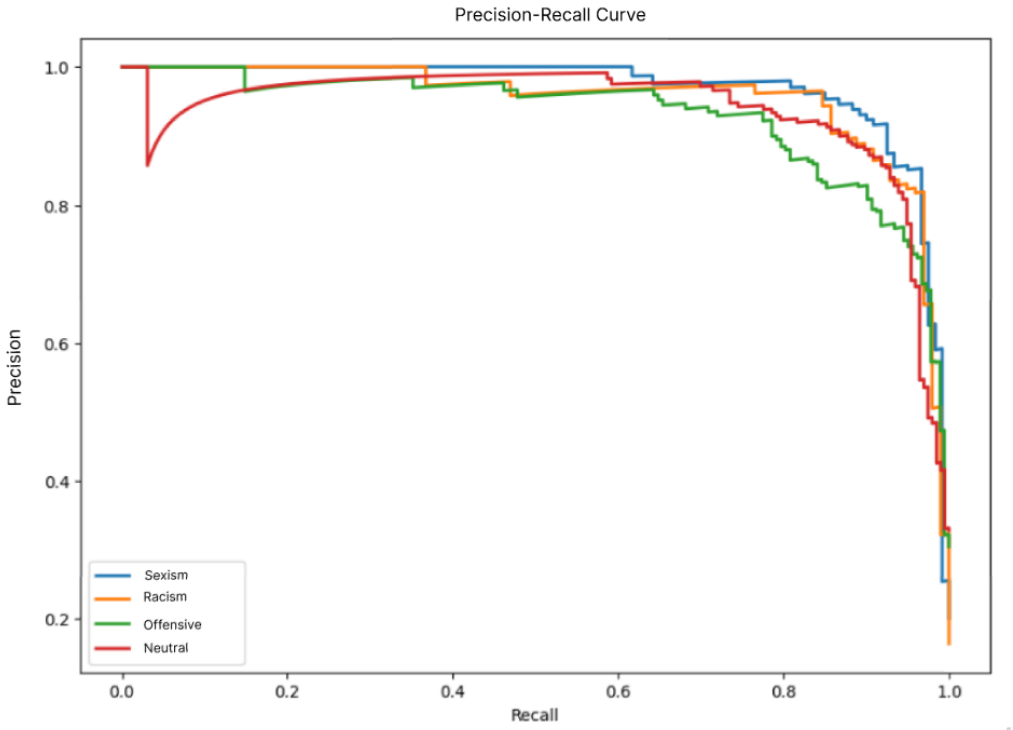

Figure 10. BERTurk PR curves

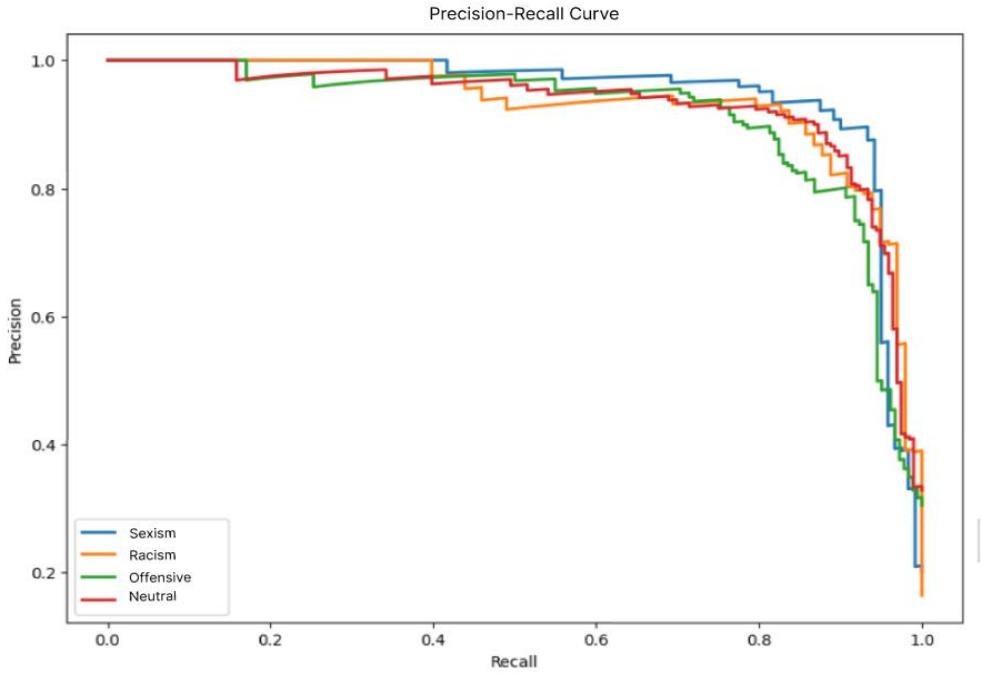

Figure 11. ConvBERTurk PR curves

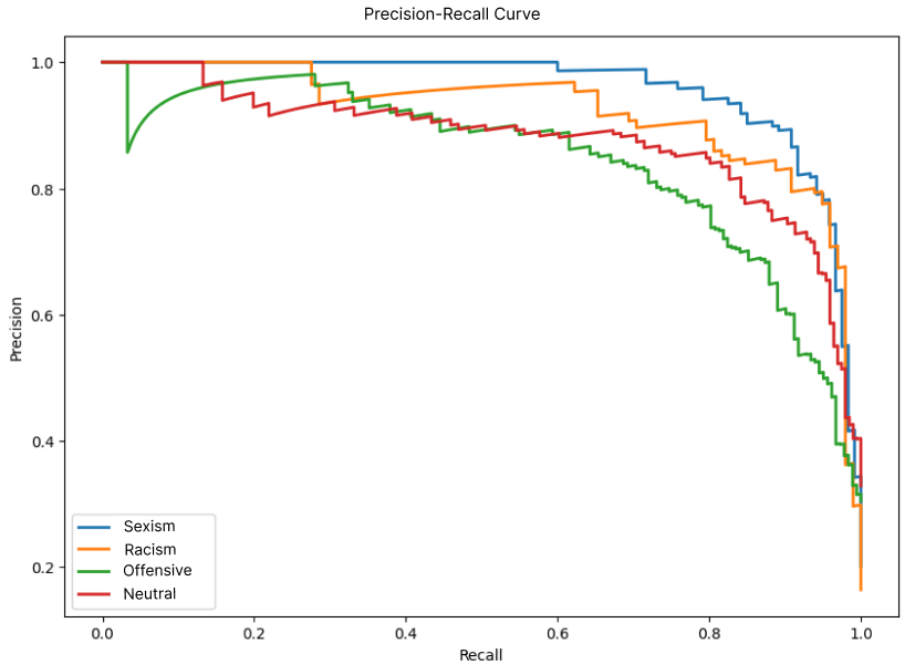

Figure 12. DistilBERTurk PR curves

The BERTurk model achieves high performance; however, the ConvBERTurk model provides similar results while offering advantages such as a shorter training duration and a smaller model file size.

5. CONCLUSIONS

Cyberbullying poses a significant threat to both the present and future world. To address this issue, raising awareness and automating processes are essential. Recent advancements in natural language processing techniques enable the categorization of text in a language by learning contextual information from large corpora. Turkish natural language processing is still developing in this field, and there is a need for using large language models in detecting cyberbullying. The developed cyberbullying detection system can be adapted for individual use, social media environments, internal corporate communication platforms, and various other settings. This contributes to raising awareness on the issue and advancing Turkish natural language processing research.

In the conducted study, several challenges were encountered. The first challenge is the scarcity of multi-class and labeled datasets in the domain of cyberbullying. The second challenge pertains to the interpretability of the language. For instance, during the testing phase, the sentence "sen de ben de ne dediğimizi bilmiyoruz" (both you and I do not know what we are saying) could be classified as either Neutral or Offensive Language by different models. Both labels could be considered correct for this sentence. Another issue is that the labeled data for the Offensive language class might be insufficiently representative of the general scope. The third issue is that in the Turkish-social-media-offensive-bullying dataset, sentences containing racial terms classified as Racism can also produce Racism outputs in contexts where they might be considered Neutral. The interpretability of the language complicates the detection of cyberbullying. Higher accuracy could be achieved with a more comprehensive and less biased Turkish dataset. The results indicate that BERT emerged as the most successful model. BERT performs well in producing strong and accurate results but operates more slowly. ConvBERT, on the other hand, provides a better balance between power and speed compared to BERT. The DistilBERT model, being a more distilled version, achieved lower accuracy compared to other models. It may be preferred in scenarios where speed and memory management are more critical. The ConvBERT model, with its balanced architecture, is suitable for both cases. Additionally, to enable users to experiment with the trained models and obtain results, as well as to analyze the models with current data, a Flask-based, user-friendly, portable, and platform-independent application was created. This API, easily integrated with software, was developed using FastAPI. Furthermore, the models can be accessed for development and testing through the Hugging Face platform [28]. The developed system provides support for those interested in model development and usage. With further development, this work could offer effective and realistic solutions for real-time detection of cyberbullying.

REFERENCES

[1] O. Zorbaz, **"Lise Öğrencilerinin Problemli İnternet Kullanımının Sosyal Kaygı ve Akran İlişkileri Açısından İncelenmesi."** Yüksek lisans tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara, 2013.

[2] F. Gültekin, **"Saldırganlık ve Öfkeyi Azaltma Programının İlköğretim İkinci Kademe Öğrencilerinin Saldırganlık ve Öfke Düzeyleri Üzerindeki Etkisi"**, Doktora Tezi, Hacettepe Üniversitesi, 2008

[3] M. Tuncer, M. Dikmen, **"Sosyal Ağlarda Bekleyen Yeni Tehlike: Siber Zorbalık"**, 4. International Instructional Technologies and Teacher Education Symposium, 94-104, 2016.

[4] İ. Yıldırım, "Sosyal Medya, Dijital Bağımlılık ve Siber Zorbalık Ekseninde Değişen Aile İlişkileri Üzerine Bir Değerlendirme" . *Anemon Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi*, 9.5: 1237-1258, 2021.

[5] E. V. Altay, B. Alataş, **"Detection of Cyberbullying in Social Networks Using Machine Learning Methods"** International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT). IEEE, p. 87-91, 3-4 Dec. 2018.

[6] V. Balakrishnan, S. Khan, H. R. Arabnia, "Improving Cyberbullying Detection Using Twitter Users' Psychological Features and Machine Learning*.", Computers & Security* 90, 101710, 2020.

[7] O. Sevli, & S. Sezgin, "Sosyal Medya Paylaşımlarında Siber Zorbalığın Tespiti ve Kategorizasyonuna Yönelik Makine Öğrenmesine Dayalı Bir Sınıflandırma". **Bursa 3rd International Scientific Research Congress,** Bursa, 626-637, 2022.

[8] D. S. Rohini, M. Ramchander, "A Comparative Study of Machine Learning Approaches for Cyberbullying Detection in Digital Forums", **International Conference on Advances in Computation, Communication and Information Technology** (ICAICCIT) (pp. 332-338). IEEE, 23-24 Nov. 2023.

[9] A. Bozyiğit, S. Utku, E. Nasibov, "Cyberbullying Detection: Utilizing Social Media Features", *Expert Systems with Applications*, 179, 115001, 2021.

[10] Ç. Çöltekin, "A Corpus of Turkish Offensive Language on Social Media**." In Proceedings of the Twelfth Language Resources and Evaluation Conference** (pp. 6174-6184). Marseille, 11–16 May 2020

[11] İ. Sel, İlhami, D. Hanbay. "Ön Eğitimli Dil Modelleri Kullanarak Türkçe Tweetlerden Cinsiyet Tespiti" *Fırat Üniversitesi Mühendislik Bilimleri Dergisi,* 33.2: 675-684, 2021.

[12] G. Nergiz, E. Avaroğlu. "Türkçe Sosyal Medya Yorumlarındaki Siber Zorbalığın Derin Öğrenme ile Tespiti." *Avrupa Bilim ve Teknoloji Dergisi* 31 :77-84, 2021.

[13] E. Karaman, "Suriyeli Mültecilere Uygulanan Ayrımcı Dışlayıcı Twitlerin BERT Modeli ile Sınıflandırılması". *Ortadoğu Ve Göç*, 12(2), 428-456, 2022.

[14] F. Beyhan, B. Çarık, I. Arın, A. Terzioğlu, B. Yanıkoğlu, & R. A. Yeniterzi, Turkish Hate Speech Dataset and Detection System. **In Proceedings of the Thirteenth Language Resources and Evaluation Conference** (pp. 4177-4185). (2022, June).

[15] A. Çelıkten, H. Bulut "Turkish Medical Text Classification Using Bert." , **29th Signal Processing and Communications Applications Conference** (SIU). IEEE, 9-11 June 2021.

[16] B. Aytan, C. O. Sakar. "Comparison of Transformer-based Models Trained in Turkish and Different Languages on Turkish Natural Language Processing Problems." **30th Signal Processing and Communications Applications Conference** (SIU). IEEE, 15-18 May 2022.

[17] M. Özkan, G. Kar, "Türkçe Dilinde Yazılan Bilimsel Metinlerin Derin Öğrenme Tekniği Uygulanarak Çoklu Sınıflandırılması*". Mühendislik Bilimleri ve Tasarım Dergisi*, 10.2: 504-519, 2022.

[18] M. Arzu, M. Aydoğan, "Türkçe Duygu Sınıflandırma İçin Transformers Tabanlı Mimarilerin Karşılaştırılmalı Analizi", *Computer Science*, (IDAP-2023), 1-6, 2023

[19] Internet: Nanelimon, Huggingface Datasets, https://huggingface.co/datasets/nanelimon/turkish-social-media-offensive-dataset, 1.03.2024.

[20] Internet: A Corpus of Turkish Offensive Language, https://coltekin.github.io/offensive-turkish, 16.10.2024.

[21] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding" , arXiv preprint arXiv:1810.04805, 2018.

[22] S. K. Behera, R. Dash, "A Novel Feature Selection Technique for Enhancing the Performance of Unbalanced Text Classification Problem". *Intelligent Decision Technologies*, 16(1), 51-69, 2022.

[23] Z. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, S. Yan, "Convbert: Improving Bert with Span-Based Dynamic Convolution." Advances in Neural Information Processing Systems, 33: 12837-12848, 2020.

[24] V. Sanh, L. Debut, J. Chaumond, T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter", arXiv preprint arXiv:1910.01108, 2019.

[25] T. Wolf, L. Debut, V. Sanh, J.Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. V. Platen, C. Ma, Y.Jernite, Julien Plu, C. Xu, T. L. Scao, S. Gugger, M.Drame, Q. Lhoest, A., "Rush, Transformers: State-of-the-art Natural Language Processing"., **Conference on Empirical Methods in Natural Language Processing: System Demonstrations** (pp. 38-45), October 2020.

[26] M. Turan, "**Derin Öğrenme ile Beklenti Tabanlı Duygu Analizi**", Yüksek Lisans Tezi, Bursa Uludağ Üniversitesi, Fen Bilimleri Enstitüsü, 2022.

[27] H. A. Ardaç, P. Erdoğmuş, "Question-Answering System with Text Mining and Deep Networks". Evolving Systems, 1-13, 2024.

[28] İnternet: B. N. Bekar, HuggingFace, https://huggingface.co/AIZinu, 21.7.2024.