Cyberbullying Detection

Yavuz Çetin, Osman Berkay Sukas
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{yavuz.cetin1, berkay.sukas}@std.yildiz.edu.tr

***Özetçe* —Bu projenin amacı sosyal medya mesajları üzerinde doğal dil işleme teknikleri uygulanarak siber zorbalık tespiti yapmaktır. Twitter üzerinden toplanmış olan bir veri seti üzerinde beş farklı siber zorbalık türüne göre sınıflandırma yapılarak farklı vektörleştirme methodları ve farklı makine öğrenmesi ve derin öğrenme teknikleri karşılaştırılmıştır.**

Anahtar Kelimeler—Makine Öğrenmesi, Derin Öğrenme, Siber Zorbalık, Ön İşleme, Naive Bayes, SVM, KNN, LSTM, BERT, Lojistik Regresyon, Doğal Dil İşleme

***Abstract*—The aim of this project is to detect cyberbullying by applying natural language processing techniques on social media messages. Different vectorisation methods and different machine learning and deep learning techniques are compared by classifying five different types of cyberbullying on a dataset collected from Twitter.**

Keywords—Machine Learning, Deep Learning, Cyberbullying, Preprocessing, Naive Bayes, SVM, KNN, LSTM, BERT, Logistic Regression, Natural Language Processing

I. Introduction

With the developing technology, there has been a significant increase in the amount of internet usage of people. Due to various disasters such as pandemics and earthquakes in recent years, people's use of social media has reached an all-time high. These increases have brought with them various negative consequences. One of these can be given as cyberbullying activities that have increased in recent years.

Cyberbullying is a form of bullying on the internet that includes harmful and unpleasant behaviours towards other users by malicious people. Examples of cyberbullying include sending insulting messages, posting harassing comments or pictures, using someone's personal information without permission, and creating fake accounts to cause discomfort to others. In addition, posting embarrassing photos of someone or manipulating online content to make the victim look bad can also be considered within the scope of cyberbullying [1].

Cyberbullying can be done in 8 different ways on social media [2].

- **Humiliation:** To humiliate a person by spreading unfounded rumours in order to damage the person's reputation.
- **Exclusion:** Consciously excluding someone from online groups.
- **Provocation:** Encouraging people to argue with rude and offensive messages.
- **Cheating:** Obtaining personal information or photographs by deceiving a person and sharing them on social media.
- **Harassment:** Repeatedly sending offensive messages or posting such messages online.
- **Identity Theft:** By creating fake profiles or hijacking accounts Posting content that brings a person into disrepute.
- **Surfing:** Sharing photos or personal information of a person without his/her knowledge.
- **Cyber Talk:** Sending threatening messages.

Although many social media sites take steps to prevent cyberbullying, it is seen that these efforts do not yield sufficient results. Malicious people can use various methods to avoid cyberbullying measures. To give examples of these methods, changing one or more letters in insulting messages or impersonating people's relatives are commonly used tactics.

Individuals who are subjected to cyberbullying often experience isolation from society, loss of reputation, disruption of family and business life, depression, and consequently, suicidal thoughts [3].

II. Related Works

Recently, with the increase in cyberbullying offences, there has been an increase in the number of studies on the detection of cyberbullying. In this section, information about the studies on cyberbullying detection in the literature is presented.

Çilem and Tuncay [1] evaluate the performance of a GPT-3 classification model for the task of classifying tweets from Twitter into those containing cyberbullying and those not containing cyberbullying. The model was first trained and tested with Turkish tweets, and an overall accuracy of 55% was obtained. When the data was translated into English and the model was retrained and tested, the accuracy increased to 66%. The precision, recall and F1 scores for both tweet classes were calculated as 0.65, 0.68 and 0.67 for non-cyberbullying tweets and 0.67, 0.64 and 0.65 for cyberbullying tweets, respectively. The confusion matrix of the model showed that 17 tweets correctly included cyberbullying and 9 tweets incorrectly included cyberbullying.

Ferhat et al. [4] examine machine learning techniques to detect and prevent cyberbullying, evaluate the performance of machine and deep learning models and analyse

the factors affecting this performance. In addition, the importance of data preprocessing, classification, feature extraction and selection processes in cyberbullying detection is discussed. Thus, researchers working on cyberbullying detection and prevention are provided with a general perspective on feature extraction approaches, feature selection techniques and classifier selection. In addition, it is emphasised that cyberbullying detection studies can be carried out on photo, video and audio data as well as textual data.

Onur and Sezan [5] conducted a cyberbullying detection and classification study in six different categories on a dataset consisting of approximately 48 thousand tweets. The tweets, which were made analysable by natural language processing, were classified with three different algorithms: K-Nearest Neighbour, Support Vector Machine and Random Forest. With these algorithms, 76%, 83% and 82% accuracy values were obtained respectively and it was concluded that the most successful algorithm was Support Vector Machine.

In their study, Enver and Muhammet [3] use natural language processing methods to detect cyberbullying using a ready-made Turkish dataset of 3000 sentences obtained from the sharing site called Kaggle and manually generated. Due to the novelty of the dataset used and the fact that such a large number of algorithms have not been tested in the literature before, this study is thought to make significant contributions to the literature. In this study, Bagging, Boosting, C4.5, Gradient Boosting, K-Means, KNN, LR, NB, ANN, RO, SVM, Stochastic Gradient Descent and XGBoost algorithms were used comparatively on the dataset for the first time.

Gözde and Erdinç [2] conducted cyberbullying detection on Turkish comments collected from Twitter, Instagram and YouTube social networks, which are widely used today. Classification models were created using deep learning based word embedding models and success rates were compared. FastText model showed the highest performance with 93.15% success rate. Social media comments were successfully classified using FastText model and LSTM neural network.

Ishtyaq et al. [6] analyse five different machine learning models, including LightGBM, XGBoost, Logistic Regression, Random Forest and AdaBoost, to detect cyberbullying using tweet dataset based on textual features. More than 47,000 tweets in our dataset are categorised into six classes. When we analysed the machine learning models, we observed that LightGBM significantly outperformed the other models, achieving 85.5% accuracy, 84% precision, 85% recall and 84.49% F-1 score.

Yasmine et al. [7] propose an innovative approach for fine-grained cyberbullying classification by integrating Neutrosophic Logic into a Multi-Layer Perceptron (MLP) model. The proposed model improves the classification of cyberbullying types by alleviating the difficulties arising from their ambiguity and overlapping boundaries. The incorporation of Neutrosophic Logic aims to address the uncertainty, ambiguity and instability in classification decisions and offers a more comprehensive and flexible approach to tackle complex classification scenarios. The model captures the complex relationships between cyberbullying types using the one-against-one strategy in MLP classification due to overlaps and ambiguous examples. The testing phase of this model emphasises the importance of Neutrosophic Logic by using class probabilities obtained from multiple one-against-one classifiers to provide a comprehensive view of the classification results. The results of the proposed model demonstrate the performance improvement of integrating Neutrosophic Logic in fine-grained cyberbullying classification tasks.

Arwa et al. [8] deal with the challenge of automatically identifying cyberbullying in tweets from a publicly available cyberbullying dataset. This work uses a robustly optimised approach of bidirectional encoder representations from transformers (RoBERTa) using global vectors (GloVe) word embedding features. The proposed approach is compared with state-of-the-art machine, deep learning and transformer-based learning approaches using FastText word embedding. Statistical results show that the proposed model outperforms the others and achieves 95% accuracy in detecting cyberbullying tweets. Moreover, the model achieved 95% precision, 97% recall and 96% F1 score. k-fold cross-validation results further confirm the superiority of the proposed model with an average accuracy of 95.07%.

III. Natural Language Processing

Natural Language Processing is a field of artificial intelligence that enables machines to understand and process human languages [5]. For Natural Language Processing applications, texts need to be converted into a format that can be processed by computers. This transformation ensures that these texts can be understood by computers while preserving their semantic meaning. The basic stages of this process are as follows:

- **Removal of unnecessary expressions:** Numbers, hashtags, emojis, emoticons, punctuation marks, usernames, links are removed from the text data.
- **Parsing into words:** Texts are divided first into sentences and then into words. This is usually done using the space character or other separators.
- **Lemmatization:** By morphological analysis of the words, inflectional suffixes are removed, for example, the words ‘word’ and ‘words’ are converted into the same root word ‘word’. In this process, meaningful root parts of words are obtained.
- **Removal of stopwords:** These are words that do not usually make a major contribution to the meaning of a sentence, but are used frequently. These words can cause noise which, if not removed, can have a negative impact on the interpretation process.
- **Word representations:** In order for computers to process texts mathematically, mathematical representations of spoken expressions are created. In this process, word representations are obtained using various techniques.
    - **Countvectorizer:** Text documents are retrieved and vectors containing the number

of words in each document are created. These vectors can be given as input to a machine learning model.

- **TF-IDF:** By combining the frequency of a term in a document with its frequency of occurrence in all documents, it determines the importance of that term in the document. TF-IDF is an effective tool to identify important terms in a document and to distinguish documents from each other.

IV. Data Set

The data set used is Cyberbullying Classification [9]. The data set used consists of 47,692 English twitter messages. The types of cyberbullying found in the data set are shown in Table 1.

Table 1 Cyberbullying Types

Cyberbullying Types	Count
Religion	7998
Age	7992
Gender	7973
Ethnicity	7961
Not Cyberbullying	7945
Other Cyberbullying	7823

As a result of the analyses conducted throughout this research, it was concluded that the content of the other cyberbullying class could not be considered as cyberbullying. Therefore, this class was not used throughout the research.

Within the scope of this research, in addition to the five class classification studies, the data set was also classified according to whether there is cyberbullying or not. For this reason, the data set used for binary classification purpose was smaller than the one with the five classes. It can be seen in Table 2.

Table 2 Binary Set

Cyberbullying Types	Count
Not Cyberbullying	7945
Cyberbullying	6400

The refined data set, focusing on distinct categories of cyberbullying, allowed for a detailed analysis of specific types such as religion, age, gender, and ethnicity. However, the exclusion of the "Other Cyberbullying" category ensured that the research remained focused on well-defined instances of cyberbullying, enhancing the clarity and reliability of the findings. This approach also facilitated the creation of a binary classification system, distinguishing between cyberbullying and non-cyberbullying content, which simplifies the detection process in practical applications. Consequently, the research not only provides insights into the prevalence and characteristics of different types of cyberbullying but also offers a streamlined method for identifying harmful content in a broad range of contexts.

The 30 most frequently used words in the data set are given in Figure 1.

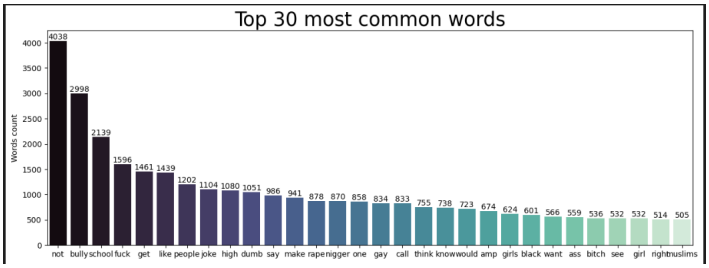

Figure 1 Most Commonly Used 30 Words

The word clouds of each cyberbullying class can be seen in Figures 2, 3, 4 and 5.

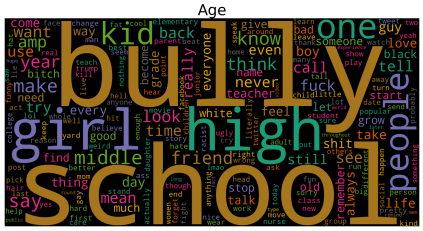

Figure 2 Age Word Cloud

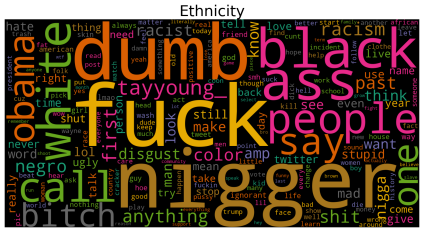

Figure 3 Ethnicity Word Cloud

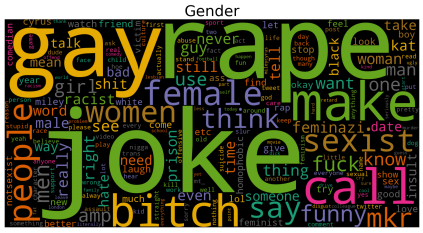

Figure 4 Gender Word Cloud

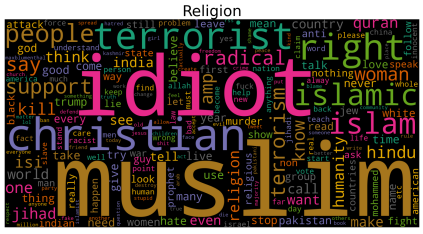

Figure 5 Religion Word Cloud

V. Machine Learning

Machine learning is a branch of artificial intelligence dedicated to creating systems capable of learning from data and making decisions accordingly. It utilizes algorithms and statistical models to enable computers to execute particular tasks without being explicitly programmed. Instead, these systems depend on recognizing patterns and drawing inferences from the data. In this section, machine learning algorithms that were used in scope of this research will be explained.

A. K Nearest Neighbors (KNN)

This classification algorithm is one of the algorithms frequently used in various artificial intelligence projects. The reason why it is frequently preferred is that it does not require training and works very fast. It uses the similarity ratios in the data to classify. It can also be explained as the process of classifying the data at a point by looking at how the data around it are classified. This process can be likened to a selection process. Given K points, all the closest points add 1 vote to their class. Thus, the class to be classified is classified into the class with the most votes. Euclidean Relation is usually used to find the neighbourhood distance of the data used. This artificial intelligence algorithm method is suitable for cyberbullying projects because it does not require training and has advantages against noise [3].

Working logic of K Nearest Neighbors can be seen in Figure 6.

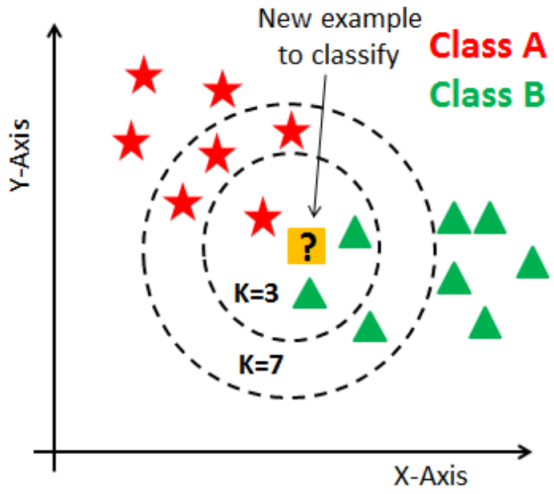

Figure 6 K Nearest Neighbors [10]

B. Logistic Regression (LR)

This classification algorithm is used to predict a binary outcome given different variables. Logistic regression divides the probability of a job or event as realisable or not realisable. The equation used in logistic regression is obtained by substituting a sigmoid function into linear regression. It is frequently used in cyberbullying detection and classification projects involving more than one class, for the separation of classes and establishing relationships between them [3].

Working logic of Logistic Regression can be seen in Figure 7.

Figure 7 Logistic Regression [11]

C. Multinomial Naive Bayes (MNB)

This classification algorithm assumes that the inputs of the classes in the data set are not interdependent and uses Bayes' Theorem to determine the effect of features on classification. For each class, the probabilities of the features are calculated and the most probable class is selected using these probabilities. This algorithm usually works fast because it considers the classes independent of each other. However, if the expected results of the inputs in the data set are related to each other, this assumption may lead to poor performance of the model. Therefore, in some cases, the desired result may not be obtained. This classification algorithm is frequently used in cyberbullying projects because its working principle is simple and generally gives better results than other algorithms [12].

Working logic of Naive Bayes can be seen in Figure 8.

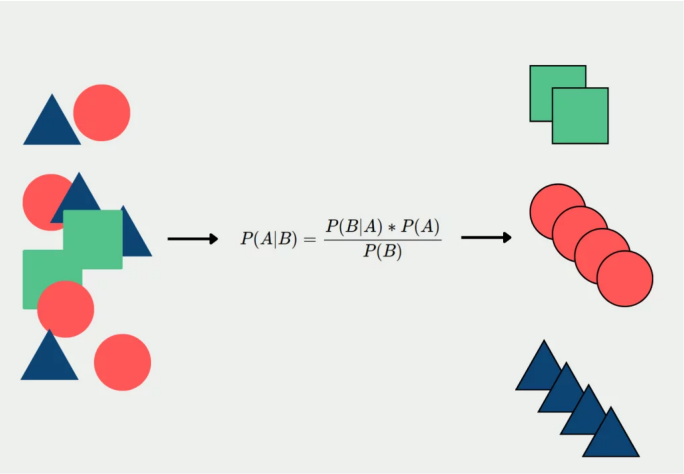

Figure 8 Naive Bayes [13]

D. Support Vector Machines (SVM)

This algorithm is one of the most frequently used algorithms in classification processes. It tries to find the best discriminative hyperplane to classify data points.While finding this hyperplane, it is aimed that the distance between two classes (edge margin) is as large as possible. This algorithm is frequently preferred in cyberbullying research due to its ability to work very fast in large data sets, to separate data linearly and non-linearly, and to select the best one among these distinctions [12].

Working logic of Support Vector Machines can be seen in Figure 9.

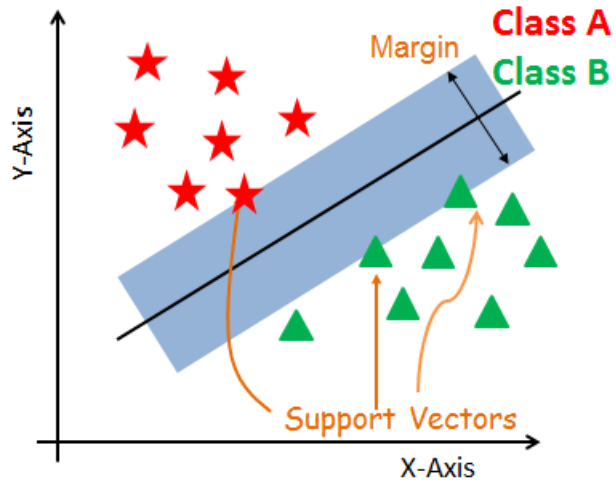

Figure 9 Support Vector Machines [14]

VI. Deep Learning

Deep learning, a branch of machine learning and artificial intelligence, focuses on algorithms that mimic the brain's structure and function, particularly through artificial neural networks. It employs multi-layered neural networks to discern and interpret complex patterns within large datasets. In deep learning models, each layer extracts increasingly abstract features from the raw data, enabling the system to effectively learn and perform tasks such as image and speech recognition, natural language processing, and autonomous driving with impressive precision. This approach has transformed numerous fields, allowing systems to enhance their performance autonomously by leveraging vast data quantities and advances in computational capabilities.

A. Long Short-Term Memory (LSTM)

It is a type of Recurrent Neural Network (RNN) architecture used in deep learning. In standard RNNs, only one hidden state information is used throughout the process. However, this prevents the system from connecting with long-term information and providing healthier results. LSTM systems use a structure called memory cell to prevent this. This structure has the ability to hold information for a long time. This structure is controlled by 3 gates. These gates are Input gate, Forget gate and Output gate.

- Input Gate - A gate that updates the state of the memory cell. As a result of giving the previous and current information to the sigmoid function, it is decided whether the memory cell will be updated or not. If the information is 0, it is considered unimportant and if it is 1, it is considered important.Here, the outputs obtained from the tanh function are multiplied by the outputs of the sigmoid function to decide which information to update [15].
- Forget Gate - Here it is decided which information is to be forgotten or stored. The decision is made after the previous information and the current information are put into the sigmoid function. Information with 0 is forgotten and information with 1 remains in the memory cell.
- Output Gate - This gate determines the input information of the next cell. First the old information and the current information is given to the sigmoid function. Then the information in the memory cell passes through the tanh function. The 2 results obtained are multiplied and thus the decision is made. As a result of these operations, hidden state information is decided.

An example of a basic LSTM cell is expressed in Figure 10.

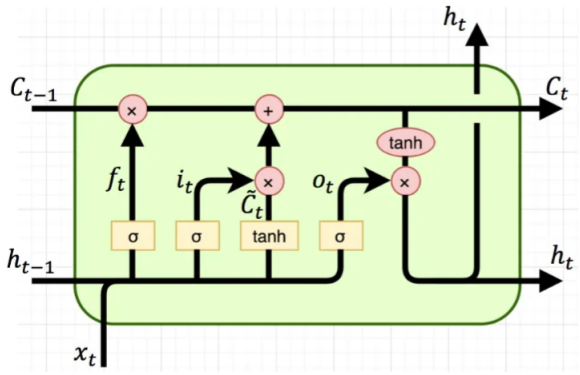

Figure 10 Basic LSTM Cell [16]

When giving embedding matrix to the LSTM model we have tried 3 different methods. These are Word2Vec, FastText and BERT embeddings.

1) Word2Vec: It is a technique frequently used in natural language processing. This model tries to model words with similar meaning as close as possible to each other numerically. Its main purpose is to represent words as high-dimensional vectors. Word2Vec typically uses one of two different approaches, CBOW (Continuous Bag of Words) and Skip-gram. CBOW tries to predict the words around a given word, while Skip-gram tries to predict the words around a given word. This model is usually trained on large text data and is used to represent word relationships in vectorial space.

2) FastText: It was developed by Facebook AI Research and is quite similar to Word2Vec, but there are some important differences. The most obvious difference is that it takes into account subwords. For example, for the word "which", it considers many sub-parts such as "wh", "whi", "hic". This is particularly useful for dealing with rare or never seen words. Also, FastText combines the vector of the word itself and the vectors of its sub-parts when calculating the vector representation of a word.

3) BERT Embeddings: These are vectors obtained from the BERT (Bidirectional Encoder Representations from Transformers) model. BERT is a Transformer-based language model and is trained unsupervised on a large text data set. One of the remarkable features of BERT is that it takes into account the context of a word to determine its meaning. Thus, the meaning of a word can be better determined by considering its relationship with other words in the text. BERT embeddings are constructed by taking this context into account and are usually obtained from pre-trained models for use in natural language processing tasks. These vectors are generally considered more effective than other embedding methods used to represent text data.

B. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a deep learning language model developed to improve the efficiency of natural language processing tasks [17]. BERT helps to understand the meaning of ambiguous language in text using the surrounding text. It is pre-trained with Wikipedia texts and can be fine-tuned with question and answer datasets.

The basis of BERT is transformers, a deep learning model in which each output element is linked to each input element and their weights are dynamically calculated. BERT has a bidirectional structure that can read texts in both directions (left to right and right to left) simultaneously. This feature is made possible by the introduction of transformer models.

The BERT model has two training processes [18].

1) Pre-training: BERT was pre-trained with two different but related NLP tasks, masked language modelling (MLM) and next sentence prediction (NSP). The goal of MLM is to hide a word in a sentence and have the programme predict the hidden word based on its context. The goal of NSP is to predict whether two given sentences have a logical and sequential connection or whether their relationship is random.

2) Finetuning: BERT can be fine-tuned for specific NLP tasks after pre-training. During fine-tuning, the model is trained by adding a new task-specific layer. This makes the model more effective and tailored for tasks such as sentiment analysis, question-answer and named entity recognition.

These training processes can be seen in Figure 11.

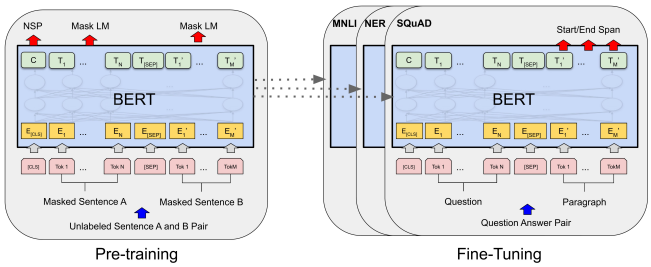

Figure 11 BERT Training Process [18]

C. A Robustly Optimized BERT Pretraining Approach (RoBERTa)

RoBERTa is a simple but very popular alternative to BERT. It primarily improves BERT's training hyperparameters by carefully and thoroughly optimising them. With a few simple and straightforward changes, RoBERTa's performance improves and allows it to outperform BERT in almost all tasks that BERT solves [19]. It uses exactly the same architecture as BERT. However, unlike BERT, the pre-training process is trained with MLM only.

D. A Lite BERT (ALBERT)

ALBERT is an alternative to BERT and is a model that produces more efficient results. Although all connections in BERT have the same size, this size is reduced in ALBERT to make the model lighter. This reduction process enables ALBERT to perform better with fewer parameters. Furthermore, ALBERT is trained using different sized attention heads. This allows ALBERT to learn more efficiently and requires less computational power.

VII. Conclusion

In this section we will give information about the results of our models.

A. Comparisons of the Data Set Versions

In this part we will show the experimental results we got after comparising the data set with the "Other Cyberbullying" class and without. To compare these two data sets we have tried 4 different machine learning methods. The comparisons can be seen in the Table 3.

Table 3 Accuracy Of the Data Sets

Veri Seti Hali	SVM	LR	KNN	MNB
Before	0.83	0.84	0.74	0.79
After	0.91	0.91	0.84	0.87

After doing the comparisons it was decided to proceed the project without the "Other Cyberbullying" class. This allowed us to get better results than other researchers who have worked on the same data set before. It is important to point out that the many messages included in "Other Cyberbullying" class were not actually cyberbullying and they were normal messages.

B. Comparisons of Machine Learning Models

In this part we will show the results of machine learning models that we have used in the project. These are K Nearest Neighbors, Support Vector Machines, Logistic Regression and Multinomial Naive Bayes.

Table 4 Accuracy Comparisons of Machine Learning Models

Vectorization Method	Category Count	SVM	LR	KNN	MNB
CV	Beş	0.94	0.94	0.82	0.88
TF-IDF	Beş	0.93	0.93	0.86	0.85
CV	İki	0.93	0.93	0.82	0.84
TF-IDF	İki	0.91	0.91	0.84	0.87

As it can be seen in Table 4 Multinomial Naive Bayes and K Nearest Neighbors fell behind on Support Vector Machines and Logistic Regression methods. LR and SVM got pretty similar results. And Countvectorizer got better results than TF-IDF method overall. It was concluded the SVM and LR models were both successful on determining the cyberbullying type when comparising 5 classes reaching to 94% accuracy. Meanwhile they fell behind 1% in 2 category classification. The reason for this is believed to be data set cutting in the beginning so both classes would have close sample counts.

C. Comparisons of Deep Learning Models

In this part we will compare the results we got from our deep learning models. We have used LSTM, BERT, RoBERTa and ALBERT for this research. For LSTM embeddings we have used 3 different vectorization methods. These methods are Word2Vec, FastText and BERT embeddings. We compared all models on both binary set and the data set with 5 classes.

Table 5 Accuracy Comparisons of Deep Learning Models

Models	Vectorization Method	Category Count	Accuracy
LSTM	BERT	5	0.91
LSTM	FastText	5	0.93
LSTM	Word2Vec	5	0.93
LSTM	BERT	2	0.93
LSTM	FastText	2	0.91
LSTM	Word2Vec	2	0.91
BERT	BERT	5	0.94
BERT	BERT	2	0.92
RoBERTa	RoBERTa	5	0.94
RoBERTa	RoBERTa	2	0.92
ALBERT	ALBERT	5	0.94
ALBERT	ALBERT	2	0.90

When the results are analysed, lower rates are observed in two-category classifications. The reason for this is that the number of samples in the data set is aimed to be the same in each class, so the data set is smaller. In the models, a data set of 47k in length was used for five-category classification, while a data set of 18k in length was used for two-category classification.

Considering the performance of all models, BERT models gave the best result with an accuracy rate of 0.94 in the five-category classification. In two categories, the best result was given by the BERT vectorisation method and LSTM with an accuracy of 0.93.

D. Comparisons with Previous Work

In this part we will compare the results of our project with previous work with the same data set. The comparison can be seen in Table 6.

Table 6 Comparison with Previous Work

Related Works	SVM	LR	KNN	MNB	BERT	RoBERTa
Our Project	0.94	0.94	0.86	0.88	0.94	0.94
Ishtyaq et al. [6]		0.83				
Yasmine et al. [7]	0.91	0.90				
Arwa et al. [8]	0.82		0.74	0.85	0.93	0.95

When the SVM model is examined, it is seen that the results obtained within the scope of this project outperform the previous studies by a rate of 3%.

When the LR model is examined, it is seen that this method is ahead with a difference of at least 4% compared to previous studies.

In the KNN model, it was observed that this project was 12% ahead of previous studies.

When MNB was used, it was observed that this project was ahead of the other study with a difference of 3%.

In the BERT model, our project was ahead of the other project by 1%, while this project was 1% behind when RoBERTa was used.

When the results obtained in general are analysed and the previous datasets using this dataset are considered, it is concluded that the results obtained within the scope of this project are more successful compared to other researches. The reason for this difference in success is thought to be that the hyper-parameters given to the models were selected through many parameter trials. It is also thought that the change of techniques applied in the pre-processing step has a significant effect on the result.

E. Best Models for Cyberbullying Classification

In this part we will select the best models to classify cyberbullying texts.

After trying many models and comparing the results for five-class classification we got the best results from the BERT model with an accuracy of 94.80%. For binary classification the best models were SVM, LR and LSTM with BERT embeddings with 93% accuracy. But its important to point out that LR model was significantly faster on this task.

References

[1] Ç. Koçak and T. YİĞİT, "Gpt-3 sınıflandırma modeli ile türkçe twitlerin siber zorbalık durumlarının belirlenmesi," *Gazi Mühendislik Bilimleri Dergisi*, vol. 9, no. 4-ICAIAME 2023, pp. 278–285, 2023.

[2] G. NERGİZ and E. AVAROĞLU, "Türkçe sosyal medya yorumlarındaki siber zorbalığın derin öğrenme ile tespiti," *Avrupa Bilim ve Teknoloji Dergisi*, no. 31, pp. 77–84, 2021.

[3] E. YAZĞILI and M. BAYKARA, "Türkçe metinlerde makine öğrenmesi yöntemleri ile siber zorbalık tespiti," *Gümüşhane Üniversitesi Fen Bilimleri Dergisi*, vol. 12, no. 2, pp. 443–453, 2022.

[4] M. F. YAKUT, Ç. ŞAHİN, and A. Yılmaz, "Machine and deep learning studies for cyberbullying detection," *Savunma Bilimleri Dergisi*, vol. 1, no. 43, pp. 155–177, 2023.

[5] O. SEVLİ and S. SEZGİN, "Sosyal medya paylaşimlarinda siber zorbaliğin tespiti ve kategorizasyonuna yönelik makine öğrenmesine dayali bir siniflandirma," *Bursa 3rd International Scientific Research Congress*, 2022.

[6] M. I. Mahmud, M. Mamun, and A. Abdelgawad, "A deep analysis of textual features based cyberbullying detection using machine learning," in *2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*. IEEE, 2022, pp. 166–170.

[7] Y. M. Ibrahim, R. Essameldin, and S. M. Saad, "Social media forensics: An adaptive cyberbullying-related hate speech detection approach based on neural networks with uncertainty," *IEEE Access*, 2024.

[8] F. Alrowais, A. A. Jamjoom, H. Karamti, M. Umer, S. Alsubai, T.-H. Kim, and I. Ashraf, "Robertanet: Enhanced roberta transformer based model for cyberbullying detection with glove features," *IEEE Access*, 2024.

[9] Larxel. (2020) Cyberbullying classification. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data

[10] R. Shah. (2021) Introduction to k-nearest neighbors (knn) algorithm. [Online]. Available: https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8

[11] Natassha. (2022) Logistic regression explained in 7 minutes. [Online]. Available: https://www.natasshaselvaraj.com/logistic-regression-explained-in-7-minutes/

[12] E. YAZĞILI and M. BAYKARA, "Siber zorbalık tespit yöntemleri potansiyel uygulama alanları ve zorluklar," *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, vol. 12, no. 1, pp. 23–35, 2021.

[13] Anonymous. (2022) What is the naive bayes algorithm? [Online]. Available: https://databasecamp.de/en/ml/naive-bayes-algorithm

[14] Hichigo. (2020) Scikit-learn ile destek vektör makineleri (svm). [Online]. Available: https://www.turkhackteam.org/konular/scikit-learn-ile-destek-vektor-makineleri-svm.1909633/

[15] O. Akköse. (2020) Uzun-kısa vadeli bellek(lstm). [Online]. Available: https://medium.com/deep-learning-turkiye/b018c07174a3

[16] K. Mani. (2019) Gru's and lstm's. [Online]. Available: https://towardsdatascience.com/grus-and-lstm-s-741709a9b9b1

[17] J. Schulze. (2024) What is the bert model and how does it work? [Online]. Available: https://www.coursera.org/articles/bert-model

[18] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[19] S. Kotamraju. (2022) Everything you need to know about albert, roberta, and distilbert. [Online]. Available: https://towardsdatascience.com/everything-you-need-to-know-about-albert-roberta-and-distilbert-11a74334b2da