



MACHINE LEARNING AND DEEP LEARNING BASED CLASSIFICATION OF PROFANITY AND SEXIST DISCOURSE IN TURKISH TEXTS

Naim KARASEKRETER¹, Caner BALIM^{2*}, Özkan ASLAN¹

¹Afyon Kocatepe University, Faculty of Engineering, Department of Computer Engineering, 03200, Afyonkarahisar, Türkiye

²Afyon Kocatepe University, Faculty of Engineering, Department of Software Engineering, 03200, Afyonkarahisar, Türkiye

Abstract: In this study, harmful content detection in Turkish texts is addressed as a three-class text classification problem comprising Harmless, Profanity and Sexist Discourse. For this purpose, open-access Turkish datasets and data compiled within the scope of the TÜBİTAK 1001 project were combined, and after removing cross-source duplicates and conflicting duplicates, a final dataset of 12051 texts was used. In the experimental process, traditional machine learning models and Transformer-based language models were compared under the same dataset, the same class structure and a 5-fold group-based stratified cross-validation protocol. For the traditional models, in addition to word- and character-level TF-IDF representations, Word2Vec- and FastText-based representations were used to evaluate Linear SVM, Logistic Regression, Decision Tree, Random Forest, Extra Trees and XGBoost algorithms. For the Transformer-based models, BERTurk, Electra-TR, ConvBERTurk and XLM-RoBERTa were adapted to the three-class classification task. Among the traditional models, the highest performance was obtained with Logistic Regression on TF-IDF representations (82.82% macro F1-score); Linear SVM produced very close results, while FastText- and Word2Vec-based representations remained comparatively lower. The obtained results showed that Transformer-based models offer higher and more balanced performance than traditional machine learning models. Among all models, the highest success was achieved with the ConvBERTurk model, which reached a 90.93% macro F1-score, 90.93% accuracy and a 98.09% ROC-AUC value. The BERTurk model exhibited a performance very close to ConvBERTurk with a 90.28% macro F1-score. A paired t-test over the five folds shows that this difference is not statistically significant ($t(4) = 2.36, P=0.078$). These findings indicate that contextual language representations are effective in Turkish harmful content classification. The study contributes to the literature by addressing Turkish harmful content detection as a multi-class problem that also includes the distinction between profanity and sexist discourse, and by comparing traditional models and Transformer-based models under the same leakage-controlled experimental protocol.

Keywords: Turkish text classification, Harmful content detection, Profanity detection, Sexist discourse detection, Machine learning, Transformer models

*Corresponding author: Afyon Kocatepe University, Faculty of Engineering, Department of Software Engineering, 03200, Afyonkarahisar, Türkiye

E mail: cbalim@aku.edu.tr (C. BALIM)

Naim KARASEKRETER  <https://orcid.org/0000-0003-2892-6430>

Caner BALIM  <https://orcid.org/0000-0002-1010-129X>

ÖZKAN ASLAN  <https://orcid.org/0000-0002-2680-5419>

Received: July 17, 2026

Accepted: August 23, 2026

Published: September 15, 2026

Cite as: Karasekreter, N., Balim C., & Aslan Ö. (2026). Machine learning and deep learning based classification of profanity and sexist discourse in turkish texts. *Black Sea Journal of Engineering and Science*, 9(5): 2676-2691.

1. Introduction

With the proliferation of digital communication environments, social media platforms, forums, news comments and online discussion spaces have become fundamental communication channels where individuals can quickly share their thoughts. Although these environments offer important opportunities for information sharing and social interaction, they also create fertile ground for the spread of harmful texts such as profanity, insults, degrading discourse and sexist discourse. As user-generated content reaches large volumes, detecting such texts solely through human moderation becomes inadequate in terms of time, cost and scalability.

In online texts, not only the presence but also the type of harmful content is important. A text may directly contain

profane or vulgar expressions, or it may contain sexist discourse — degrading, exclusionary or hateful expressions targeting a person or group on the basis of gender. In contrast, texts that do not contain aggression or inappropriateness can be regarded as harmless. In this study, Turkish texts are addressed within the scope of a three-class classification problem consisting of Harmless, Profanity and Sexist Discourse.

Harmful content detection in Turkish texts involves various difficulties due to the structural characteristics of the language and online writing habits. While the agglutinative structure of Turkish causes words derived from the same root to be used in many forms with different suffixes, the spelling errors, abbreviations, character repetitions, emojis, usernames, hashtags and irregular punctuation marks seen in social media texts



make automatic text processing more complex.

The main aim of this study is to comparatively examine the performance of models that can automatically distinguish these three classes. Accordingly, traditional machine learning models and Transformer-based language models were tested under the same dataset and evaluation protocol. While word- and character-level TF-IDF representations were used in the traditional models, pre-trained contextual language representations were utilized in the Transformer-based models. Thus, the performance difference between classical feature-extraction-based methods and modern deep-learning-based approaches in Turkish harmful content classification was systematically analyzed. The contribution of this study can be evaluated at two levels. At the data level, the original labels obtained from different sources were not transferred directly to the three-class structure; they were mapped onto a consistent three-class scheme within the framework of the annotation guideline and the inclusion-exclusion criteria defined by the researchers, and label groups that did not match the class definitions were excluded from the scope. At the method level, by applying a group-based stratified cross-validation protocol that prevents conflicting duplicates and identical samples at the normalized-text level from leaking together into the training and test partitions, traditional machine learning models and four different Transformer-based models were systematically compared under the same experimental conditions. In this respect, the contribution of the study is not so much to propose a new model architecture as to bring together a multi-source three-class dataset for Turkish harmful content detection under a leakage-controlled, reproducible evaluation protocol, enabling different model families to be evaluated under comparable conditions.

1.2. Literature Review

Esen and Amrahov (2024) reported 0.858 accuracy, 0.856 F1-score and 0.923 ROC-AUC with a Bernoulli Naive Bayes model on 12989 Turkish texts. Yakar et al. (2024) compared classical algorithms with feature-selection methods on two different Turkish tweet datasets; they reported an F-measure of 88.81% with OneR + Naive Bayes on the 1K dataset and 87.71% with Info-Gain + Naive Bayes on the 2K dataset. Similarly, Kesgin and Demirer (2025) compared Random Forest, Naive Bayes, Logistic Regression, SVM and LSTM models for offensive language detection on 11253 posts collected from Twitter, Ekşi Sözlük and forums; they reported 0.82 accuracy and 0.65 AUC for the Random Forest model without Word2Vec, and 0.78 accuracy and 0.79 AUC for the LSTM model. Aliyeva and Yağanoğlu (2025) proposed an MLP-based model that jointly uses textual and social-media features for cyberbullying detection in Turkish Twitter posts; with TF-IDF + unigram features they reported 93.2 accuracy, 91.22 precision, 95.6 sensitivity and 93.36 F1-score. In addition to classical machine learning methods, representation-learning and deep-

learning approaches are widely used in Turkish harmful content detection. Karayığıt (2026) compared classical machine learning and deep learning models on a dataset of Instagram comments for detecting insult content that constitutes a crime under the Turkish Penal Code; the TF-IDF + Naive Bayes model obtained 0.85 macro-averaged F1, while the BERT embedding + attention-based CNN model obtained 0.93 macro-averaged F1. Gümüş et al. (2024) used transformer-based sentence vectors with classical classifiers on the Turkish offensive language dataset of Tanyel et al. (2022) and reported 0.8512 F1 with the Multilingual-MiniLM + SVM combination. Najafi and Varol (2024) used more than 894 million Turkish tweets from 2010–2020 in their TurkishBERTtweet model developed on Twitter/X data and adapted the RoBERTa-base architecture with pre-processing steps specific to social-media language. The model achieved 0.807 weighted F1 on the HateSpeech SIU dataset, 0.815 weighted F1 on the HSD2LANG dataset and 0.73167 macro F1 in the official SIU 2023 split for hate-speech detection tasks.

There are also studies comparing BERT-based models with large language models in Turkish hate-speech detection. Dehghan and Yanıkoğlu (2024) compared ChatGPT with a fine-tuned BERTurk model for Turkish hate-speech detection on the extended SIU2023-NST dataset; BERTurk-CE achieved 82.22% accuracy and a 74.86% macro-F1 score, while ChatGPT achieved 70.81% accuracy and a 58.50% macro-F1 score in the zero-shot setting and 65.81% accuracy and a 60.19% macro-F1 score in the few-shot setting. Aksoy et al. (2026) reported 0.958 precision, 0.879 recall and 0.916 F1-score for soft voting in 10-fold cross-validation with a transformer-based ensemble approach for detecting hate speech directed at LGBT individuals.

Studies conducted within the scope of shared tasks have enabled the comparison of different transformer architectures in Turkish hate-speech detection. Within the HSD-2Lang 2024 shared task, Singhal and Bedi (2024) fine-tuned XLM-RoBERTa Large for hate-speech detection in Turkish tweets and reported an F1 score of 0.63258 in the Turkish subtask. In the same task, Yagci et al. (2024) fine-tuned the TurkishBERTtweet model without using additional data; they obtained 0.74 F1 on the public test set and 0.69 F1 on the private test set in the Turkish subtask, ranking second on the private leaderboard. Barkhordar et al. (2024) used the dbmdz/bert-base-turkish-128k-uncased model; they reported 0.71365 F1 on the public test set and 0.60790 F1 on the private test set in the official submission, and 0.71889 and 0.66129 F1 respectively in the late submission.

Data imbalance and data-augmentation approaches are also considered among the important factors affecting model success in Turkish harmful content detection. Karayığıt et al. (2024) compared different balancing techniques on the ATC dataset; with TextGAN-based balancing they reported Macro-Averaged F1 values of

0.972 for CNN models, 0.958 for SVM, 0.954 for Random Forest and 0.955 for Naive Bayes. Mut Altın and Saggion (2025) examined sexist discourse and its subtypes on 6915 Turkish texts collected from X and YouTube; with GPT-4o-mini fine-tuning they reported 0.91 F1 in binary classification and 0.59 F1 in multi-class sexism-type classification. De Oliveira et al. (2024) evaluated the GPT-3 Ada model with multilingual corpora and a cross-lingual learning strategy in the political-discourse domain; by adding a Turkish political tweet corpus among the source languages, they obtained high F1-scores in different target languages. It should be noted that Turkish was used only as a source language in that study and was not evaluated as a target language.

Studies in the field of harmful content detection are diversifying beyond text-level classification toward target detection, token-level labeling, hybrid architectures and multimodal applications. Uludoğan et al. (2024) used the TurkishHatePrintCorpus dataset of 6681 news articles for hate-speech detection in the Turkish print media and proposed the HateTargetBERT approach, which combines BERTurk representations with target-oriented linguistic features. The model achieved 90.54% accuracy, 88.47% precision, 90.82% recall and 89.60% F1-score on the test set. Büyükdemirci et al. (2024) addressed text-level neutral/offensive/hateful classification on the JL-Hate dataset together with token-level target and expression-span labels; for Turkish they reported 71.7% Macro F1 in text classification and 52.3% Macro F1 in token classification. Altinel et al. (2024) combined BERT representations with a k-means and textGCN-based hybrid approach; they reported F-measures of 77.43%, 77.81% and 81.17% on the 1K, 2K and 9K datasets respectively, and an 87.81% F1-score in the 9K scenario where all features are used together.

In general, studies in the literature show a trend from traditional machine learning methods toward Transformer-based models in Turkish harmful content detection; on the other hand, problems such as data imbalance, lack of context, semantic proximity between classes and the discrimination of harmful-content subtypes persist. Taking these trends in the current literature into account, this study addresses Turkish texts in three classes - Harmless, Profanity and Sexist Discourse - and compares traditional machine learning models with Transformer-based models under the same experimental protocol. A comparative summary of the studies reviewed in this section is presented in Table 1.

2. Materials and Methods

In this study, a three-class text classification approach was applied to automatically detect harmful content in Turkish texts. In this context, the aim was to determine which of the Harmless, Profanity or Sexist Discourse classes each text belongs to. In the experimental design, traditional machine learning algorithms and Transformer-based language models were compared

using the same dataset, the same cross-validation strategy and the same evaluation metrics. Thus, the performance of classical approaches based on TF-IDF word- and character-level text representations and modern models leveraging contextual language representations such as BERT, RoBERTa, Electra and ConvBERT in Turkish harmful content classification was examined comparatively and systematically.

2.1. Dataset

The dataset used in this study was created to address harmful content detection in Turkish texts as a three-class classification problem. The dataset was prepared by combining open-access Turkish datasets with texts compiled within the scope of the TÜBİTAK 1001 project numbered 124E056. The data sources used consist of the Turkish offensive language corpus published by Çöltekin (2020), the Turkish toxic language dataset provided by Overfit-GM (2023), the gender subset of the large-scale multi-domain hate speech corpus published by Toraman et al. (2022a,b), and additional texts compiled as part of the project. The data obtained from different sources were reorganized to fit the class structure of this study. In this context, each text was assigned to one of three classes: Harmless, Profanity and Sexist Discourse. The Harmless class refers to texts that do not contain aggression, profanity or sexist degradation. The Profanity class covers expressions containing rude, profane and obscene language. The Sexist Discourse class covers degrading, exclusionary or hateful expressions targeting a person or group on the basis of gender or gender identity.

The final dataset consists of a total of 12051 samples: 4580 samples taken from the Overfit-GM dataset, 4977 samples from the gender subset of the corpus of Toraman et al. (2022a,b) 2080 samples from the Çöltekin corpus and 414 samples compiled within the scope of the TÜBİTAK 1001 project. The labels in the source datasets are based on the original labels produced by human annotators in the annotation processes of the respective studies. These labels were not transferred directly and as-is to the three-class structure used in this study; inclusion and exclusion criteria were defined for each class in line with the annotation guideline created by the researchers, and the source labels were mapped onto the three-class scheme according to these criteria. The texts compiled within the TÜBİTAK 1001 project were labelled by three scholarship-holding annotators, all of whom are graduates of undergraduate programmes related to Turkish language, using the same annotation guideline applied to the external corpora. Each text was labelled independently by all three annotators; the final label was determined by majority agreement and the remaining disagreements were resolved through joint discussion. Of these texts, only those assigned to the Harmless and Profanity classes were transferred to the final dataset (400 and 14 samples respectively), while texts labelled as obscene content were excluded because they do not correspond to any of the three classes.

Table 1. Summary comparison of studies in the field of Turkish harmful content detection

Year	Reference	Dataset	Class Distribution	Methodology	F1	Accuracy	AUC	Precision	Recall
2024	Esen and Amrahov	Private	No: 6850, Yes: 6139, Total: 12989	Bernoulli Naive Bayes	0.856	0.858	0.923	0.863	0.858
2024	Yakar et al.	Public	None: 1287, Offensive expression: 387, Hate speech: 375, Total: 2049	Feature Extraction + ML Algorithms	0.8771	-	-	-	-
2024	Gümüş et al.	Public	Offensive: Not specified, Non-offensive: Not specified, Total: 53005 (42398 train / 8851 test / 1756 validation)	Sentence Transformer + ML Models	0.8512	-	-	0.8515	0.8512
2024	Najafi and Varol	Public	No hate speech: 6121, Hate speech: 2684, Total: 8805	TurkishBERTweet (RoBERTa-base)	0.815	-	-	-	-
2024	Dehghan and Yanıkoğlu	Public	No hate speech: 5596, Hate speech: 1808, Total: 7404	BERTurk + ChatGPT zero-shot / few-shot	0.7486	0.8222	-	0.7612	0.7389
2024	Singhal and Bedi	Public	Non-hateful: 6258, Hateful: 2882, Total: 9140	XLM-RoBERTa Large	0.63258	-	-	-	-
2024	Yagci et al.	Public	Non-hateful: 6258, Hateful: 2882, Total: 9140	TurkishBERTweet	0.74	-	-	-	-
2024	Barkhordar et al.	Public	Non-hateful: 6258, Hateful: 2882, Total: 9140	BERTurk	0.71365	-	-	-	-
2024	Karayığit et al.	Private	Non-abusive: 19826, Abusive: 10528, Total: 30354	TextGAN + CNN + SVM + RF + Naive Bayes	0.972	-	0.98	0.98	0.91
2024	De Oliveira et al.	Public	Turkish corpus used as source language: No hate speech: 5055, Hate speech: 2512, Total: 7567	GPT-3 Ada (target language: English; Turkish used as a source language)	0.9658	-	-	0.9689	0.9695
2024	Uludoğan et al.	Public	Non-hateful: 3678, Hateful: 3003, Total: 6681	HateTargetBERT	0.8960	0.9054	-	0.8847	0.9082
2024	Büyükdemirci et al.	Public	Neutral: 258, Offensive: 349, Hateful: 67, Skipped: 68, Tie: 23, Total: 765	Transformer Models	0.717	-	-	-	-
2024	Altinel et al.	Public	None: 6447, Hate speech: 1757, Offensive expression: 1614, Total: 9818	Transformer Models + TextGCN	0.8781	-	-	-	-
2025	Kesgin and Demirer	Private	Offensive: Not specified, Non-offensive: Not specified, Total: 11253	TF-IDF + ML Models (Random Forest, without Word2Vec)	-	0.82	0.65	-	-
2025	Aliyeva and Yağanoğlu	Public	Non-cyberbullying: 2500, Cyberbullying: 2500, Total: 5000	MLP + TF-IDF/Unigram + Social Media Features	0.9336	0.932	0.932	0.9122	0.956
2025	Mut Altın and Saggion	Public	Not-sexist: 3167, Sexual violence: 1352, Stereotyping: 1124, Misogyny: 655, Objectification: 468, Anti-feminism: 149, Total: 6915	GPT-4o mini Fine-Tuning	0.91	-	-	-	-
2026	Karayığit	Public	Non-insult: 3709, Insult: 3652, Total: 7361	Feature Extraction + ML and DL	0.93	0.93	-	0.93	0.93
2026	Aksoy et al.	Public	Hate speech: 4247, Non-hate speech: 4247, Total: 8494	Transformers + Soft Voting	0.916	0.9201	-	0.958	0.879

In this context, the non- and prof-labeled samples in the Çöltekin corpus were mapped to the Harmless and Profanity classes respectively; the OTHER-, PROFANITY- and SEXIST-labeled samples in the Overfit-GM dataset were mapped to the Harmless, Profanity and Sexist Discourse classes respectively; and the normal- and hate-labeled samples in the corpus of Toraman et al. (2022a,b) were mapped to the Harmless and Sexist Discourse classes. The samples from the TÜBİTAK project were included in the dataset with their harmless and profanity labels preserved. Label groups that did not match the class definitions — the INSULT and RACIST labels in the Overfit-GM dataset, the general insult categories directed at individuals or groups in the Çöltekin corpus, the offensive label with blurred class boundaries in the Toraman corpus, machine-translated content and non-Turkish lines — were excluded from the scope. Thus, the Sexist Discourse class was formed from the union of the Overfit-GM SEXIST-labeled samples (670 samples) and the hate-labeled samples of the Toraman corpus (2779 samples).

During the data-merging process, attention was paid to preserving the semantic distinction between classes; near-duplicates across sources and conflicting-duplicate groups — in which the same text, or texts considered identical in normalized form, appeared with different labels — were removed from the dataset. In addition, by including the harmless-labeled samples from the gender-themed source in the Harmless class, the aim was for the models to learn actual sexist discourse rather than the mere presence of the gender topic, thereby reducing the risk of topic-based shortcut learning. To measure the level of leakage that could arise from the source-label structure, a baseline evaluation using only the source information was performed; the accuracy of this baseline was calculated as 0.567 and fell within the pre-specified acceptance range (0.50–0.60) (majority-class baseline: 0.436). The fact that the source-label association cannot be completely eliminated — since sexist content comes predominantly from gender-focused sources and profanity content predominantly from profanity-focused sources — was documented as a structural limitation of the study.

The group variable used in the experiments was also produced with the same normalization approach. For this purpose, the texts were passed through Unicode normalization, converted to lowercase while preserving the Turkish character structure; URLs, usernames and numerical expressions were replaced with standard tokens, punctuation marks were separated with spaces and excess whitespace was cleaned. As a result of this process, samples with the same normalized text were grouped under the same group identifier. In this study, the grouping operation is based on the exact matching of normalized texts rather than a character-based approximate-similarity threshold. Thus, identical texts — or those considered identical in normalized form — were prevented from being distributed to different cross-validation folds.

As a result of these operations, a duplicate-free final dataset of 12051 texts was used in the experiments. In the final dataset, 5255 texts are in the Harmless class, 3347 texts in the Profanity class and 3449 texts in the Sexist Discourse class (Table 2). Thus, a consistent and near-balanced experimental data structure was obtained that allows traditional machine learning models and Transformer-based models to be compared under the same dataset and the same class structure.

The classes are unevenly distributed across the four sources (Table 3). In particular, 2779 of the 3449 Sexist Discourse texts (80.6%) come from the gender subset of Toraman et al. (2022a,b) and the remaining 670 (19.4%) from the Overfit-GM dataset, while 73.3% of the Profanity class originates from Overfit-GM. Together with the source-only baseline reported above (0.567 accuracy), this concentration is addressed as a limitation in Section 6.

Table 2. Class distribution of the final dataset

Class	Label	Count
Harmless	0	5255
Profanity	1	3347
Sexist Discourse	2	3449
Total	-	12051

Table 3. Distribution of the final dataset across sources and classes

Source	Harmless	Profanity	Sexist Discourse	Total
Toraman et al. (2022a,b) – gender subset	2198	0	2779 (80.6%)	4977
Overfit-GM (2023)	1457	2453	670 (19.4%)	4580
Çöltekin (2020) – Troff	1200	880	0	2080
TÜBİTAK 1001 project (124E056)	400	14	0	414
Total	5255	3347	3449	12051

2.2. Data Preprocessing

In the data-preprocessing stage, the aim was to make the texts in the dataset suitable for model training. First, basic normalization operations were applied to the texts.

In this context, all texts were standardized in Unicode form, excess whitespace was cleaned, and records containing empty text were removed from the dataset. In addition, it was checked that the label values consisted

only of the values 0, 1 and 2 representing the three classes. This operation was carried out to prevent problems that could arise from erroneous or unexpected labels during the classification process.

In the preprocessing process, different text-preparation approaches were used for traditional machine learning models and Transformer-based models. In the traditional machine learning models, the aim was to reduce superficial noise that could negatively affect the classification process while preserving the semantic content of the texts. In this context, the texts were converted to lowercase taking the Turkish character structure into account; URLs, usernames, hashtag expressions, emojis and numerical values were removed from the texts. In addition, repeated punctuation marks that are common in social media texts were simplified, and consecutively repeated characters were standardized. As a result of these adjustments, the texts were brought into a more consistent form, and the aim was to extract TF-IDF-based word and character features more stably.

For the Transformer-based models, a more limited preprocessing method was preferred. In these models, only Unicode normalization and whitespace-adjustment operations were applied to the texts. Elements such as punctuation marks, capitalization and emojis were preserved. This approach was adopted to enable the Transformer-based models to benefit from the contextual and superficial cues in the text. Thus, while more standardized texts were used for the traditional models, the original structure of the text was largely preserved in the Transformer models.

2.3. Experimental Design

The experimental design is based on comparing traditional machine learning models with Transformer-based language models in Turkish harmful content classification. All models were tested on the same dataset using the same class structure and the same evaluation setup. Thus, the aim was to ensure that the differences between model performances stem from the representation method and model architecture used, rather than from data splitting or different evaluation conditions. While hyperparameter search with inner cross-validation was performed in the traditional models, standard fine-tuning values were used in the Transformer-based models; this difference in the tuning approach is detailed in Section 3.3.2. The overall experimental workflow is summarized in Figure 1.

To measure model performance more reliably and to reduce the risk of data leakage, 5-fold group-based stratified cross-validation was applied in the experiments. For this operation, the StratifiedGroupKFold approach was used; the number of folds was set to 5, the shuffle option was active and the random-state value was set to 42. The class label was used as the stratification variable, and the duplicate-

group key produced from the normalized texts was used as the group variable. Thus, the aim was to preserve the class distribution as much as possible in each fold and to prevent samples with the same normalized text from leaking together into the training and test partitions.

Before cross-validation, it was checked whether each group contained more than one different class label. Cases in which the same normalized text appeared with different labels were regarded as conflicting-duplicate groups, and all samples in these groups were removed from the dataset. In the fold-creation stage, it was also verified that each class had sufficient samples and sufficient different groups to enable 5-fold evaluation. After these checks, all models were evaluated on the same outer cross-validation folds.

In the traditional machine learning models, feature extraction and classifier training were performed within each outer cross-validation fold using only the training data. In the TF-IDF-based experiments, the word- and character-level vectorizers were defined within a Pipeline structure together with the classifier; thus, the TF-IDF vocabulary, IDF weights and model parameters were learned only on the relevant training fold. The test fold was transformed and evaluated only through the trained pipeline. This approach prevented the test data from being included in the feature-extraction or hyperparameter-selection process.

In the traditional models where hyperparameter selection was performed, the search was carried out only with inner cross-validation run on the training portion of the outer cross-validation. In the inner cross-validation, group-based stratified splitting was used where possible, and the macro F1-score was taken as the selection criterion. Thus, the outer test fold was not used in the model-selection, feature-extraction or hyperparameter-tuning stages.

Accuracy, macro F1-score, macro precision, macro recall and multi-class ROC-AUC were used as evaluation metrics. In this study, the macro F1-score was taken as the main comparison metric. The macro F1-score was obtained by taking the arithmetic mean of the F1 values calculated separately for each class; thus, the aim was for each class to contribute equally to model performance regardless of class sizes. Similarly, precision and recall values were also calculated as macro averages.

The multi-class ROC-AUC value was calculated with a one-vs-rest approach in which each class is evaluated separately against the other classes, and the macro average was taken across classes. In the Transformer-based models, the class probabilities obtained with softmax were used in the ROC-AUC calculation. In the traditional machine learning models, predict_proba was used for models that provide a probability output, and class scores were used for models that provide a decision-function output.

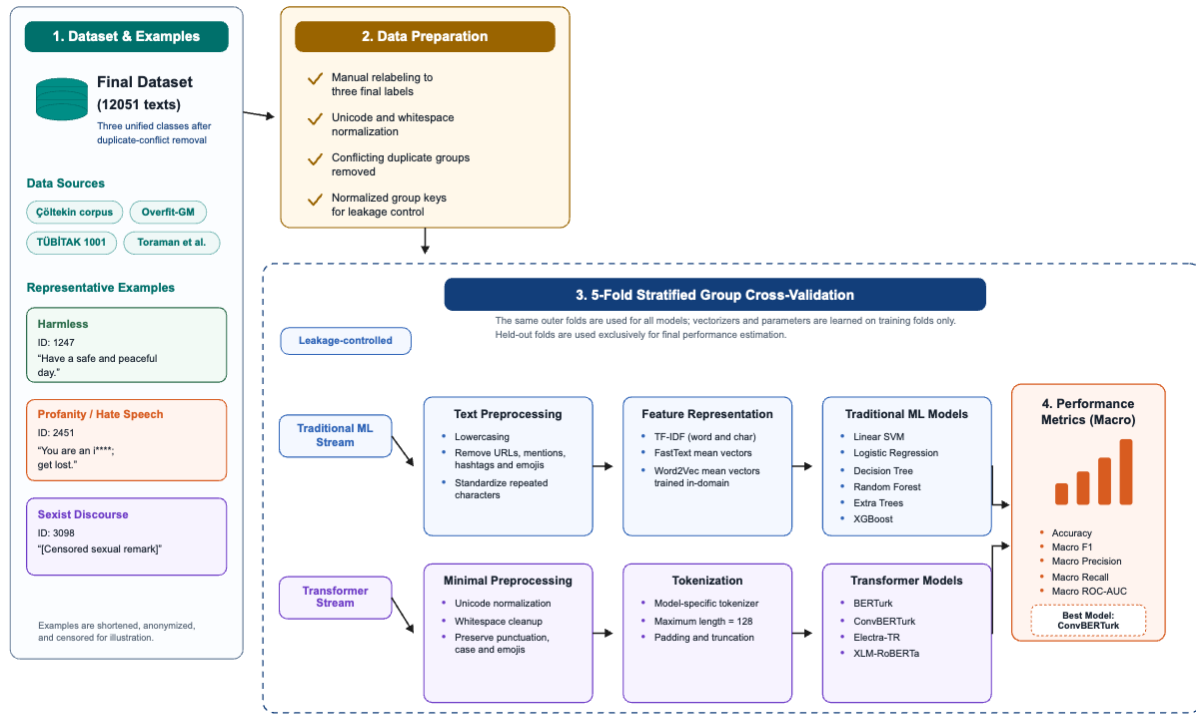


Figure 1. Experimental workflow diagram for Turkish harmful content classification.

2.3.1. Machine learning models

In the traditional machine learning models, the texts were numerically represented using TF-IDF-based feature extraction, which incorporates inverse document frequency weighting (Sparck Jones, 1972). In this process, both word-level and character-level n-gram features were used. At the word level, unigrams and bigrams were considered, while at the character level, three-, four- and five-gram character patterns were used. The joint use of word and character representations aims to reflect to the model both the meaningful word patterns in the text and the spelling variations and character-level formal cues that are common in Turkish social media texts (Table 4).

Table 4. Hyperparameter settings for TF-IDF feature extraction

Hyperparameter	Value
Word TF-IDF n-gram	(1, 2)
Word max features	50000
Character TF-IDF Analyzer	char_wb
Character n-gram	(3, 5)
Character Max Features	75000
Use IDF	True
Smooth IDF	True
Padding	True
Norm	L2 Norm

In the Word2Vec experiments, the embedding model was retrained for each outer cross-validation fold using only

the relevant training fold. The texts in the test fold were represented using the Word2Vec (Mikolov et al., 2013) vocabulary learned on the training fold; words not present in the training vocabulary were not included in the vector average. Thus, Word2Vec representations were prevented from learning information from the test data. In the FastText experiments, no retraining was performed on the dataset; pre-trained Turkish cc.tr.300 FastText vectors (Bojanowski et al., 2017; Grave et al., 2018) were used. FastText-based text representations were computed separately for the training and test folds; the classifiers were trained only on the training fold (Table 5).

Table 5. Hyperparameter settings for Word2Vec feature extraction

Hyperparameter	Value
Representation	Mean Vector
Vector Size	300
Epoch	10
Window	5
Min Count	1
Skip-gram	1

Linear Support Vector Machine (Cortes and Vapnik, 1995), Logistic Regression (Cox, 1958), Decision Tree (Breiman et al., 1984), Random Forest (Breiman, 2001), Extra Trees (Geurts et al., 2006) and XGBoost (Chen and Guestrin, 2016) algorithms were evaluated on these features. In the Logistic Regression and Linear Support Vector Machine models, the regularization parameter

was tried over different values. In the Decision Tree model, different model configurations were tested by varying the splitting criterion and the minimum number of samples required at a leaf node. In the Random Forest and Extra Trees models, hyperparameter search was performed over the number of trees and the minimum leaf-sample values. Hyperparameter selection was carried out within the training folds, and the most suitable model configuration was determined according to the macro F1-score.

2.3.2. Transformer models

In the Transformer-based experiments, four different language models that can leverage Turkish and multilingual text representations were used: BERTurk, Electra-TR and ConvBERTurk, which are Turkish pre-trained models based respectively on the BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020) and ConvBERT (Jiang et al., 2020) architectures, and XLM-RoBERTa (Conneau et al., 2020), which extends the RoBERTa pre-training approach (Liu et al., 2019) to a multilingual setting. The publicly released checkpoints dbmdz/bert-base-turkish-cased, dbmdz/electra-base-turkish-cased-discriminator, dbmdz/convbert-base-turkish-cased and xlm-roberta-base were used respectively. These models were adapted to the three-class text classification task and trained separately for each cross-validation fold. The model output layer was arranged to represent the Harmless, Profanity and Sexist Discourse classes.

In the Transformer-based models, too, each outer cross-validation fold was treated as an independent experiment. In each fold, the relevant pre-trained model was reloaded from its initial weights, fine-tuning was performed only on the training fold, and the test fold was used only for the final evaluation. Since the tokenizer operation was performed with the fixed vocabulary of the pre-trained model, no additional vocabulary learning or parameter update was performed on the test data.

In the Transformer models, the texts were tokenized and the maximum sequence length was set to 128. The training process was run for 3 epochs, with a training batch size of 16 and an evaluation batch size of 32. The learning rate was set to 2×10^{-5} and the weight-decay coefficient to 0.01. Each model was fine-tuned on the relevant training fold and then evaluated on the test portion of the same fold. The logit values obtained in the model outputs were converted into probability values, and the class with the highest probability was accepted as the model's prediction.

Unlike the traditional models, no model-specific hyperparameter search was performed in the Transformer-based models; instead, standard fine-tuning values widely adopted in the literature (learning rate 2×10^{-5} , weight decay 0.01, 3 epochs, sequence length 128) were applied uniformly to all four models. This choice was adopted both to limit computational cost and to ensure that the difference between models stems from architectural capacity rather than from the tuning budget. Therefore, the Transformer models were

evaluated without any tuning advantage compared with the traditional models (Table 6). Since this asymmetry is in favour of the traditional models, the performance gap reported in Section 4 should be read as a conservative lower bound on the Transformer advantage.

Table 6. Hyperparameter settings used for the Transformer models

Hyperparameter	Value
Max Sequence Length	128
Epoch	3
Train Batch Size	16
Validation Batch Size	32
Learning Rate	2×10^{-5}
Weight Decay	0.01
Random Seed	42
Padding	True
Truncation	True
Optimizer	AdamW
Warmup Steps	0
Scheduler	Linear

In the Transformer-based models, each outer cross-validation fold was run as an independent experiment. In each fold, the relevant pre-trained model was reloaded from its initial weights and fine-tuned only on the training fold for 3 epochs. The test fold was used only after training was completed, to obtain the final predictions and to compute the metrics. Therefore, the samples in the test fold were not included in the model-selection or parameter-update process.

3. Results

This section presents the experimental results of traditional machine learning models and Transformer-based language models on Turkish harmful content classification. In the experiments, each model was evaluated with 5-fold cross-validation, and F1-score, accuracy, ROC-AUC, precision and recall values were computed for each fold. Then, comparative results were obtained by taking the fold averages of these metrics for each model. The results were examined under two main groups: traditional machine learning models and Transformer-based models.

3.1. Machine Learning Results

In the traditional machine learning models, Logistic Regression, Linear SVM, Decision Tree, Random Forest, Extra Trees and XGBoost models were evaluated using TF-IDF-based word and character features. According to the fold averages, the highest F1-score among the traditional models was obtained by the Logistic Regression model with 82.82%. Detailed results for all models are presented in Table 7.

Among the traditional machine learning models, linear

models — particularly the Linear SVM and Logistic Regression approaches — produced more balanced results with TF-IDF-based word and character features. The markedly lower metrics of the Decision Tree model compared with the other models can be explained by the structural limitations of the single decision-tree structure with respect to the text representations used. Since decision trees make axis-aligned decisions over a single feature at each split, they can capture discriminative class patterns in the tens-of-thousands-dimensional and sparse TF-IDF space only with very deep tree structures that are prone to overfitting. In dense embedding representations such as Word2Vec and FastText, where class information is distributed across many dimensions,

a single embedding dimension is not discriminative on its own, so axis-aligned splits cannot represent the cross-dimensional linear combinations that linear models learn easily. In addition, the high-variance structure of unpruned single trees limits generalization performance; indeed, the fact that the Random Forest and Extra Trees models, which use the same base learner in ensemble form, produced F1 values approximately 6–15 points higher than the Decision Tree model across all three representation types confirms that the observed low performance largely stems from the variance of the single tree. Nevertheless, these ensemble models also remained behind the F1-score and accuracy values of the linear models.

Table 7. TF-IDF-based machine learning model results

Model	F1	Accuracy	AUC	Precision	Recall
Linear SVM	82.43%	82.74%	94.43%	83.41%	81.74%
Logistic Regression	82.82%	83.11%	94.66%	83.85%	82.12%
Decision Tree	74.79%	74.91%	81.58%	74.97%	74.67%
Random Forest	81.05%	81.14%	94.10%	82.29%	80.22%
Extra Trees	82.00%	82.17%	94.37%	83.21%	81.19%
XGBoost	76.16%	76.82%	91.13%	79.65%	74.71%

Table 8 presents a comparison of the F1-score and accuracy values of the machine learning models obtained using the TF-IDF, FastText and Word2Vec feature-extraction methods. When the results are evaluated overall, it is seen that the highest performance was obtained with the Logistic Regression model on TF-IDF representations. The Logistic Regression model reached an 82.82% F1-score and 83.11% accuracy with TF-IDF features, producing the most successful result among all model and feature-extraction-method combinations. This model was followed by the Linear SVM model with an 82.43% F1-score and 82.74% accuracy, again on TF-IDF representations.

The relatively low F1 value (76.16%) produced by the XGBoost model on TF-IDF representations stems from the compatibility of the high-dimensional and sparse TF-IDF feature space with the gradient-boosted tree structure. The fact that shallow decision trees can use only a limited number of features at each split makes it difficult for the model to capture discriminative patterns in such a sparse space. Indeed, the same model reaching a 77.81% F1 value on the 300-dimensional dense FastText representations and thus exceeding its TF-IDF-based result supports this interpretation.

When FastText representations are used, the most successful model is again seen to be Linear SVM. The Linear SVM model reached a 78.35% F1-score and 78.96% accuracy with FastText features; although these results remained approximately 4.5 points behind the best TF-IDF-based results, they were clearly above the Word2Vec representations. On the FastText representations, the Logistic Regression model ranked second with a 77.98% F1-score and 78.57% accuracy,

while the XGBoost model showed a similar level of performance with a 77.81% F1-score and 78.46% accuracy. The fact that the pre-trained subword-based FastText vectors produced higher results than the Word2Vec representations demonstrates the contribution of subword information and pre-training on a large corpus in the agglutinative structure of Turkish. When the Word2Vec average-vector representations are used, the model performances remain lower than with the TF-IDF and FastText representations. In the Word2Vec-based results, the highest F1-score was obtained by the Linear SVM model with 73.76%. The Extra Trees (73.14%), Random Forest (73.00%) and XGBoost (72.97%) models produced moderate results close to each other, while the Decision Tree model showed the comparatively lowest performance on the Word2Vec representations with a 62.96% F1-score.

3.2. Transformer Model Results

Within the scope of the Transformer-based models, the BERTurk, ConvBERTurk, XLM-RoBERTa and Electra-TR models were evaluated under the same dataset and cross-validation structure. Considering the fold averages, the highest F1-score was obtained by the ConvBERTurk model with 90.93%. The ConvBERTurk model demonstrated strong and balanced performance across all evaluation metrics, with 90.93% accuracy, 90.96% precision and 90.94% recall. This model was followed by the BERTurk model with a 90.28% F1-score and 90.28% accuracy. Detailed results for all models are presented in Table 9.

Table 8. Performance comparison of machine learning models by feature-extraction method

Model	TF-IDF		FastText		Word2Vec	
	F1	ACC	F1	ACC	F1	ACC
Linear SVM	82.43%	82.74%	78.35%	78.96%	73.76%	74.08%
Logistic Regression	82.82%	83.11%	77.98%	78.57%	72.79%	72.99%
Decision Tree	74.79%	74.91%	60.84%	61.63%	62.96%	63.10%
Random Forest	81.05%	81.14%	75.65%	76.79%	73.00%	73.34%
Extra Trees	82.00%	82.17%	75.79%	76.87%	73.14%	73.50%
XGBoost	76.16%	76.82%	77.81%	78.46%	72.97%	73.21%

Table 9. Transformer model results

Model	F1 (mean $\pm$ SD)	Accuracy	AUC	Precision	Recall
ConvBERTurk	90.93% $\pm$ 0.45	90.93%	98.09%	90.96%	90.94%
BERTurk	90.28% $\pm$ 0.54	90.28%	97.95%	90.47%	90.16%
XLM-RoBERTa	87.29% $\pm$ 0.39	87.29%	96.84%	87.72%	86.99%
Electra-TR	89.94% $\pm$ 0.43	89.91%	97.74%	90.17%	89.77%

The findings in Table 9 show that the Transformer-based models reached a high level of success in the classification task. The fact that the BERTurk and ConvBERTurk models produced very close values in the F1-score, accuracy, precision and recall metrics reveals that pre-trained language models for Turkish provide stable and consistent results in this task. The Electra-TR model was in the high-performance group with an 89.94% F1-score and 89.91% accuracy. Although the XLM-RoBERTa model produced lower results than the other Transformer models, it exhibited a strong classification performance with F1-score and accuracy values above 87%.

When the macro ROC-AUC values are examined, all Transformer models are seen to reach values above 96.8%. This result shows that the models have a high capacity to separate the classes from each other within

the scope of one-vs-rest evaluation. As a result of the overall evaluation, it was determined that the Transformer-based models provide higher and more balanced performance than the traditional machine learning approaches. Among the evaluated Transformer models, the highest overall performance was obtained by the ConvBERTurk model. However, as shown in Section 4.2.1, its difference from BERTurk is not statistically significant. To examine the class-level prediction success of the ConvBERTurk model in more detail, a confusion matrix was created by combining all fold results. When the confusion matrix in Figure 2 is examined, the model achieved a high correct-classification rate in all three classes. This result shows that the ConvBERTurk model exhibits a balanced structure not only in the overall metrics but also in class-level discrimination success.

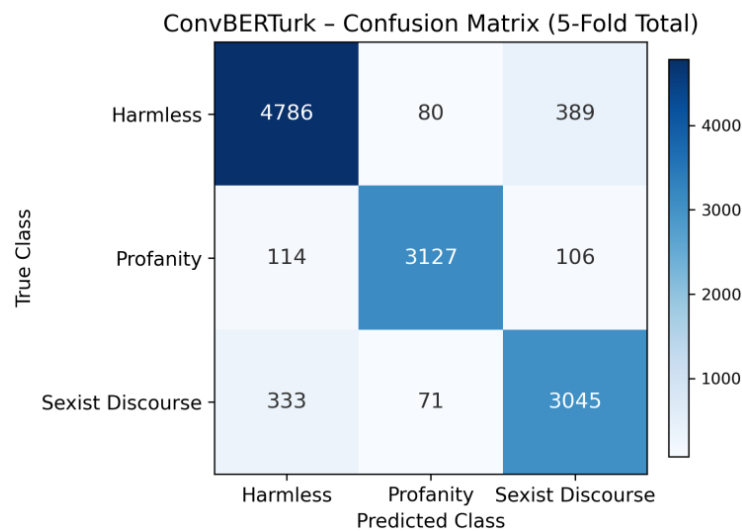


Figure 2. Confusion matrix for the ConvBERTurk model (out-of-fold predictions, all five folds).

The model is seen to reach the highest number of correct predictions in the Harmless class, correctly classifying 4786 samples belonging to this class. In the Profanity class, 3127 samples, and in the Sexist Discourse class, 3045 samples, were correctly predicted. Misclassifications are seen to concentrate particularly between the Harmless and Sexist Discourse classes; 389 Harmless samples were predicted as Sexist Discourse, and 333 Sexist Discourse samples were predicted as Harmless. The low confusion rate of the Profanity class with the other classes can be explained by the strong lexical cues carried by profane expressions. The Harmless–Sexist Discourse confusion, on the other hand, stems from the fact that sexist discourse is often expressed through context and implication without explicit profanity. Nevertheless, the low total number of errors reveals that the model makes a successful distinction between the classes.

3.2.1. Statistical significance of the differences between the transformer models

Since the differences among the top-performing models are small in absolute terms, the macro F1 values obtained

in the five folds were compared pairwise using two-sided paired t-tests and the Wilcoxon signed-rank test. The difference between ConvBERTurk (90.93% $\pm$ 0.45) and BERTurk (90.28% $\pm$ 0.54) is not statistically significant ($t(4) = 2.36$, $P=0.078$; Wilcoxon $P=0.063$), nor is the difference between BERTurk and Electra-TR (89.94% $\pm$ 0.43; $t(4) = 2.28$, $P=0.085$). By contrast, the ConvBERTurk–Electra-TR difference ($t(4) = 7.65$, $P=0.002$) and all comparisons against XLM-RoBERTa (87.29% $\pm$ 0.39; $P<0.001$) are statistically significant.

ConvBERTurk, BERTurk and Electra-TR therefore form a single top-performance group that cannot be statistically separated at the 0.05 level, and only XLM-RoBERTa falls clearly below them. The 0.65-point advantage of ConvBERTurk is accordingly reported as the highest observed mean rather than as evidence of an absolute best model, and the related statements in the Abstract, Findings and Discussion sections have been formulated in line with this qualification. It should also be noted that, with five folds, the smallest attainable two-sided p-value of the Wilcoxon test is 0.063, so the t-test results are taken as the primary evidence.

Table 10. Comparison of the obtained results with the 2025-2026 Turkish harmful content detection literature

Year	Reference	Dataset	Class Distribution	Methodology	F1	Accuracy	AUC	Precision	Recall
2026	This Publication	Public + Private	Harmless: 5255, Profanity: 3347, Sexist discourse: 3449, Total: 12051	ConvBERTurk	0.9093	0.9093	0.9809	0.9096	0.9094
2025	Kesgin and Demirer (2025)	Private	Offensive: Not specified, Non-offensive: Not specified, Total: 11253	TF-IDF + ML Models (Random Forest, without Word2Vec)	-	0.82	0.65	-	-
2025	Aliyeva and Yağanoğlu (2025)	Public	Non-cyberbullying: 2500, Cyberbullying: 2500, Total: 5000	MLP + TF-IDF/Unigram + Social Media Features	0.9336	0.932	0.932	0.9122	0.956
2025	Mut Altın and Saggion (2025)	Public	Not-sexist: 3167, Sexual violence: 1352, Stereotyping: 1124, Misogyny: 655, Objectification: 468, Anti-feminism: 149, Total: 6915	GPT-4o mini Fine-Tuning	0.91	-	-	-	-
2026	Karayığıt (2026)	Public	Non-insult: 3709, Insult: 3652, Total: 7361	Feature Extraction + ML and DL	0.93	0.93	-	0.93	0.93
2026	Aksoy et al. (2026)	Public	Hate speech: 4247, Non-hate speech: 4247, Total: 8494	Transformers + Soft Voting	0.916	0.9201	-	0.958	0.879

4. Discussion

Among the traditional machine learning models, Logistic Regression produced the highest result on TF-IDF-based representations with 82.82% F1, closely followed by Linear SVM with 82.43% F1; FastText representations produced at best 78.35% F1, while Word2Vec representations remained limited at 73.76% F1 (Tables 7-8). When moving to the Transformer-based models, a

marked performance increase was observed in all evaluation metrics (Table 9). The ConvBERTurk model reached the highest average value with a 90.93% macro F1-score, while the BERTurk model ranked second with a 90.28% macro F1-score. Since the 0.65-point difference between the two models is limited, the results show that the two models exhibit close and strong performance in this task. This can be interpreted as the ConvBERT

architecture's combination of the multi-head self-attention mechanism with span-based dynamic convolutions providing an additional advantage in detecting expressions such as profanity and sexist discourse, which are short and carry dense contextual cues. The Electra-TR model ranked third with 89.94% F1, while the multilingual XLM-RoBERTa model remained at 87.29% F1. The approximately 3.3-point average superiority of the Turkish-specific models (ConvBERTurk, BERTurk) over the multilingual model shows that learning language-specific morphological and syntactic patterns during pre-training contributes directly to task performance.

When evaluated in terms of the ROC-AUC metric, the Transformer-based models clearly exhibit a superiority over the traditional models. While the AUC values in the traditional TF-IDF-based models remained in the 81.58%–94.66% range, this value exceeded 96.8% in all four Transformer models. ConvBERTurk reached 98.09%, BERTurk 97.95%, Electra-TR 97.74% and XLM-RoBERTa 96.84% AUC values. These results show that the Transformer models can determine inter-class decision boundaries with high accuracy. The fact that ConvBERTurk produced the highest values in both the macro F1 and macro ROC-AUC metrics shows that this model also exhibits a consistent superiority in inter-class score discrimination within the scope of one-vs-rest evaluation.

When the precision and recall metrics are examined specifically for the Transformer models, ConvBERTurk is seen to offer the most balanced performance with 90.96% precision and 90.94% recall. Maintaining this balance is critically important in harmful content detection, both to keep the false-alarm rate low and to avoid missing actual harmful content. The BERTurk model also produced values quite close to ConvBERTurk, with 90.47% precision and 90.16% recall. In the XLM-RoBERTa model, precision was 87.72% and recall 86.99%; these values confirm that the multilingual model lags behind the Turkish-focused models in capturing Turkish-specific harmful discourse patterns. Considering that these metrics remained in the approximately 75%–84% range in the traditional TF-IDF-based models, the Transformer models provided an improvement of approximately 7–9 points in the precision–recall balance, and this improvement is considered to make a meaningful difference for practical applications. It is noteworthy that this superiority was obtained even though only the traditional models were additionally tuned by grid search; the reported gap is therefore a conservative lower bound on the Transformer advantage. As already discussed in Section 5 in connection with Table 10, the 0.9093 macro F1 obtained with ConvBERTurk is comparable to the closest recent Turkish results; since those values come from binary setups with different F1 definitions, the comparison remains contextual.

Within these limits, the 0.972 F1 reported by Karayiğit et

al. (2024) with TextGAN-based data balancing rests on a different experimental setup. Read against this background, the results obtained here — without any data augmentation and under a three-class formulation — indicate that ConvBERTurk provides a strong and balanced performance in Turkish harmful content detection.

To test the generalizability of the proposed experimental configuration, the same architecture and the same fine-tuning regime (ConvBERTurk, $2e-5$ learning rate, 3 epochs, group-based stratified 5-fold cross-validation) were retrained and evaluated on two Turkish datasets from the literature. On the cyberbullying dataset used by Aliyeva and Yağanoğlu (2025), a macro F1 of 93.15% (± 1.00) was obtained, which is at a level comparable to the 93.36% value reported in the related study for the MLP model based on social-media features and feature selection. However, the two values are not directly equivalent: the value reported in the related study is a positive-class F1 obtained on a single hold-out split, whereas the value obtained here is a macro F1 averaged over group-based stratified 5-fold cross-validation. On the TLHS1 dataset (Aksoy et al., 2026), a macro F1 of 0.856 (± 0.008) was obtained on three different balanced subsets. Aksoy et al. report a positive-class F1 of 0.856 for a single ConvBERT model on the same dataset; since the two values are computed with different averaging schemes, they should be read as indicating a comparable performance level rather than as a direct confirmation of the same quantity. The fact that the same model configuration produced results of 90.93% on the three-class dataset of this study, 93.15% on the cyberbullying dataset and 85.6% on TLHS1 shows that the observed performance differences stem not so much from the model or hyperparameter choice as from the task definition, labeling criteria and linguistic difficulty level of the dataset (for example, the implicit hate and irony content in TLHS1). This finding reveals that the performances reported in Turkish harmful content detection are dataset-dependent and that direct metric comparisons across studies are meaningful only through reproduction on the same dataset.

Five limitations should be taken into account when interpreting these findings. First, the class–source distribution is uneven: 80.6% of the Sexist Discourse class comes from a single corpus (Table 3), so the model's representation of this class depends largely on the annotation decisions of that corpus and should not be generalized to other platforms or annotation schemes without validation on an independently annotated external sexism corpus. Second, the hyperparameter budget is asymmetric — a grid search was applied only to the traditional models — which favours those models and makes the reported gap between the two families a conservative lower bound. Third, all texts come from social media and comparable short-form online content, so performance on longer or domain-specific texts such as news articles or forum threads was not evaluated.

Fourth, the three-class scheme excludes categories present in the source corpora, such as general insult, racist discourse and obscenity. Fifth, all experiments were run with a single random seed (42), so seed-level variability was not quantified.

5. Conclusion

In this study, TF-IDF, FastText and Word2Vec text-representation methods were systematically compared with traditional machine learning models and Transformer-based language models for the automatic detection of profanity and sexist-discourse content in Turkish texts. Among the traditional models, TF-IDF-based Logistic Regression achieved the highest performance with 82.82% F1, while Linear SVM produced very close results with 82.43% F1; FastText representations reached at best a 78.35% F1 value, while Word2Vec-based representations remained limited at 73.76% F1 (Tables 7–8). In the Transformer-based models, a marked performance jump was observed. The ConvBERTurk model exhibited the highest overall performance among all models, with a 90.93% F1-score, 90.93% accuracy and 98.09% ROC-AUC (Table 9). This superiority of ConvBERTurk can be attributed to the ConvBERT architecture's ability to model local context patterns more effectively thanks to its span-based dynamic convolution mechanism. The BERTurk model produced results very close to ConvBERTurk, with a 90.28% F1 and a 97.95% macro ROC-AUC. The Electra-TR and XLM-RoBERTa models also obtained strong results with 89.94% and 87.29% F1 respectively; however, it was determined that models trained specifically for Turkish (BERTurk, ConvBERTurk) provide a consistent superiority over multilingual models (XLM-RoBERTa). These results reveal that contextual language representations offer a clear advantage over superficial text representations in Turkish harmful content detection and that pre-trained models suited to the language structure further strengthen this advantage. However, its difference from BERTurk is not statistically significant over the five folds ($t(4) = 2.36$, $P=0.078$); ConvBERTurk, BERTurk and Electra-TR form a statistically indistinguishable top-performance group. The obtained findings reveal the position of the study more clearly when compared with recent studies published in the field of Turkish profanity, offensive language and sexist discourse in the 2024–2026 period. Since different datasets, class structures and evaluation metrics are used in the literature, a one-to-one comparison is not always possible; nevertheless, a meaningful evaluation can be made based on general trends. Table 10 presents the results of this study comparatively with selected recent studies.

The literature summary presented in Table 10 was structured to provide a contextual evaluation — rather than a direct performance race — of studies focusing on Turkish harmful content, offensive language, hate speech and cyberbullying detection in the 2025–2026 period.

Because the studies in the literature use different datasets, numbers of classes, task definitions and evaluation protocols, a one-to-one comparison of the obtained metrics involves methodological limitations. To reflect these methodological differences transparently, Table 10 was expanded to include information on the task definition, number of classes, metric type, dataset accessibility and evaluation protocol of the relevant studies. In this study, the 90.93% macro F1 and 90.93% accuracy values obtained with the ConvBERTurk model are in a success range similar to the 0.91 macro F1 reported by Mut Altın and Saggion (2025) in binary sexism classification and the 0.916 F1 reported by Aksoy et al. (2026). Obtained under a more challenging evaluation protocol — three-class, leakage-controlled and based on group-based cross-validation — these results indicate a competitive performance level with respect to recent Turkish studies within the scope of the dataset and experimental setup used in the study.

In the studies in the literature, harmful content detection is addressed within the scope of binary classification problems such as hate speech, cyberbullying, offensive language or sexist discourse, or within specific subtasks. In this study, on the other hand, the distinction between harmless content, profanity and sexist discourse in online Turkish texts was evaluated together, and the problem was modeled within a three-class classification structure. In this respect, the study offers a more comprehensive approach that addresses different types of harmful content comparatively within the same experimental framework, rather than focusing on only a single type of harmful discourse. In the obtained results, the ConvBERTurk model reached both the highest macro F1-score and accuracy values and the highest macro ROC-AUC value. This finding shows that Transformer-based models have a high inter-class discrimination capacity. At the same time, it should be taken into account that the performance differences between the ConvBERTurk and BERTurk models are quite limited; the results should be interpreted as pre-trained Transformer models for Turkish offering close and strong performance in this task, rather than as an absolute model superiority.

Nevertheless, the obtained findings should not be evaluated as a direct and absolute claim of methodological superiority over the studies in the literature. This is because the compared studies differ in terms of dataset size, data source, class distribution, labeling criteria, number of classes, evaluation protocol and the type of F1 reported. Therefore, the high performance values obtained in this study should be considered within a contextual and limited comparison framework with the results in the literature. Among the possible reasons for the high and balanced performance observed in the study are the fact that the three-class problem formulation contains discriminative lexical and contextual cues, the strong representation capacity of Transformer-based models adapted to Turkish, and the use of a group-based stratified cross-validation strategy

to reduce the risk of data leakage. In this context, the contribution of the study to the literature is not limited to quantitative performance values; the study also addresses Turkish harmful content detection within a more comprehensive classification framework, compares different model families under equal experimental conditions, and offers a systematic methodological structure that takes class balance and the risk of duplicate-induced data leakage into account in the evaluation process.

In future studies, the primary goal is to expand the 12051-text dataset used in this research with larger and more diverse data obtained from different online platforms. This will make it possible to increase the generalization success of the models against different discourse types, user profiles and content-production styles. In addition, supporting the current three-class labeling structure with more detailed subcategories and examining hierarchical classification approaches will contribute to the development of more detailed decision mechanisms in harmful content detection. Evaluating large language models such as Llama and Gemma in zero-shot, few-shot and fine-tuning scenarios; addressing different Transformer-based architectures together with ensemble-learning methods; and examining data-augmentation strategies for class imbalance are considered important for future research. These approaches are expected to increase the adaptability of the developed models to real-time content-management systems and to enable more comprehensive evaluations in terms of application-oriented metrics such as processing time, scalability and system stability. In addition, since harmful content can be produced not only in textual form but also through multimodal structures that use visual and textual elements together, addressing multimodal approaches together with explainable artificial intelligence methods will contribute to the development of more transparent, reliable and applicable harmful content detection systems in the future.

Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	N.K.	C.B.	Ö.A
C	40	30	30
D	40	30	30
S	40	30	30
DCP	40	30	30
DAI	40	30	30
L	40	30	30
W	40	30	30
CR	40	30	30
SR	40	30	30
PM	40	30	30
FA	40	30	30

C= concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

Acknowledgements

This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the 1001 Program, Project No. 124E056. The authors would like to thank all members of the project team for their valuable contributions to the compilation and labeling of the dataset. During the preparation of this work, the authors used Claude, an AI language model developed by Anthropic, to assist in drafting and refining sections of the manuscript. After using this tool, the authors reviewed and edited the content to ensure its accuracy and appropriateness. The authors take full responsibility for the content of the published article.

Data Availability / Reproducibility

With regard to reproducibility, the merged dataset will not be redistributed as raw text because the source corpora carry their own licences; instead, the sample identifiers, the source-to-three-class mapping and the dataset-construction script will be released in an open repository together with the random seed (42), the fold indices and the fold-level metric files, so that the same dataset and experiments can be reconstructed. The 414 texts compiled within the project will be made available on request within the data-sharing conditions of the project.

References

- Aksoy, Ç., Demirezen, M. U., & Sağiroğlu, Ş. (2026). Hate speech detection in Turkish: An ensemble transformer-based deep learning approach. *Engineering Applications of Artificial Intelligence*, 164, 113147. <https://doi.org/10.1016/j.engappai.2025.113147>
- Aliyeva, Ç. O., & Yağanoğlu, M. (2025). Deep learning approach to detect cyberbullying on Twitter. *Multimedia Tools and Applications*, 84(19), 20497–20520. <https://doi.org/10.1007/s11042-024-19869-3>
- Altinel, A. B., Karatas Baydogmus, G., Sahin, S., & Gurbuz, M. Z. (2024). So-haTRed: A novel hybrid system for Turkish hate speech detection in social media with ensemble deep learning improved by BERT and clustered-graph networks. *IEEE Access*, 12, 86252–86270. <https://doi.org/10.1109/ACCESS.2024.3415350>
- Barkhordar, E., Topçu, I. S., & Hürriyetoğlu, A. (2024, March). Team Curie at HSD-2Lang 2024: Hate speech detection in Turkish and Arabic tweets using BERT-based models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE 2024)* (pp. 215–220).
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. https://doi.org/10.1162/tacl_a_00051
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth International Group.
- Büyükdemirci, K., Kucukkaya, I. E., Ölmez, E., & Toraman, C. (2024). JL-Hate: An annotated dataset for joint learning of hate speech and target detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 9543–9553).
- Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In 8th International Conference on Learning Representations (ICLR 2020). <https://openreview.net/forum?id=r1xMH1BtvB>
- Çöltekin, Ç. (2020, May). A corpus of Turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6174–6184).
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). <https://doi.org/10.18653/v1/2020.acl-main.747>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Cox, D. R. (1958). The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215–242. <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>
- Dehghan, S., & Yanikoğlu, B. (2024, March). Evaluating ChatGPT's ability to detect hate speech in Turkish tweets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE 2024)* (pp. 54–59).
- De Oliveira, A. B., Baptista, C. D. S., Firmino, A. A., & De Paiva, A. C. (2024, April). A large language model approach to detect hate speech in political discourse using multiple language corpora. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing* (pp. 1461–1468). <https://doi.org/10.1145/3605098.3635964>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
- Esen, F. S., & Amrahov, S. E. (2024, October). TurkHSD: A hate speech detection model for Turkish text content. In *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1–8). IEEE. <https://doi.org/10.1109/ASYU62119.2024.10757058>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018, May). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 3483–3487). European Language Resources Association.
- Gümüş, E., Sarıyar, R., & Göz, F. (2024, September). Comparative analysis of transformer-based vector models for hate speech in Turkish tweets. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1–5). IEEE. <https://doi.org/10.1109/IDAP64064.2024.10710825>
- Jiang, Z. H., Yu, W., Zhou, D., Chen, Y., Feng, J., & Yan, S. (2020). ConvBERT: Improving BERT with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, 33, 12837–12848.
- Karayigit, H. (2026). Leveraging attention-based deep learning models for detecting insult offenses on social media in the context of Turkish Penal Law. *Neural Computing and Applications*, 38, Article 6. <https://doi.org/10.1007/s00521-025-11807-0>
- Karayigit, H., Akdağlı, A., & Acı, Ç. İ. (2024). Performance comparison of data balancing techniques on hate speech detection in Turkish. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 30(5), 610–621. <https://doi.org/10.5505/pajes.2023.40072>
- Kesgin, B. F., & Demire, R. M. (2025). Offensive language detection in Turkish language by using NLP. *Sakarya University Journal of Science*, 29(1), 1–17. <https://doi.org/10.16984/sofenbilder.1349956>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* [Preprint]. arXiv. <https://arxiv.org/abs/1907.11692>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space [Preprint]. arXiv. <https://arxiv.org/abs/1301.3781>
- Mut Altın, L. S., & Saggion, H. (2025, September). Investigating large language models' (LLMs) capabilities for sexism detection on a low-resource language. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing—Natural Language Processing in the Generative AI Era* (pp. 771–779).

- <https://doi.org/10.26615/978-954-452-098-4-089>
- Najafi, A., & Varol, O. (2024). TurkishBERTweet: Fast and reliable large language model for social media analysis. *Expert Systems with Applications*, 255, 124737. <https://doi.org/10.1016/j.eswa.2024.124737>
- Overfit-GM. (2023). *Turkish texts for toxic language detection* [Data set]. Hugging Face. Retrieved June 17, 2026, from <https://huggingface.co/datasets/Overfit-GM/turkish-toxic-language>
- Singhal, K., & Bedi, J. (2024, March). Transformers at HSD-2Lang 2024: Hate speech detection in Arabic and Turkish tweets using BERT-based architectures. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE 2024)* (pp. 190–194).
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.
- Tanyel, T., Alkurdi, B., & Ayvaz, S. (2022). Linguistic-based data augmentation approach for offensive language detection. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 1–6). IEEE.
- Toraman, C., Şahinuç, F., & Yilmaz, E. (2022a). Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 2215–2225).
- Toraman, C., Şahinuç, F., & Yilmaz, E. (2022b). Gender hate speech Turkish — gender identity subset [Data set]. Hugging Face. <https://huggingface.co/datasets/ctoraman/gender-hate-speech-turkish>
- Uludoğan, G., Yüksel, A. E., Tunçer, Ü. C., Işık, B., Korkmaz, Y., Akar, D., & Özgür, A. (2024, March). Detecting hate speech in Turkish print media: A corpus and a hybrid approach with target-oriented linguistic knowledge. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE 2024)* (pp. 205–214).
- Yagci, U. U., Kolcak, A. E., & Iscan, E. (2024, March). ReBERT at HSD-2Lang 2024: Fine-tuning BERT with AdamW for hate speech detection in Arabic and Turkish. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE 2024)* (pp. 195–198).
- Yakar, Ö., Büyüktanır, B., Çil, A. E., & Altınel Girgin, A. B. (2024). Türkçe nefret söylemi problemi analizinde farklı sınıflandırma algoritmalarının ve özellik seçimi yöntemlerinin performans karşılaştırması. *Avrupa Bilim ve Teknoloji Dergisi*, (53), 97–111.