electronics

MDPI

Article

Explainable Cyberbullying Detection for Multilingual Online Platforms Using NLP and Augmentation-Enhanced Deep Learning

Meghal Mandot, Olusola Oluwakemi Abayomi-Alli, Nikolaj Goranin * and Adebayo Adewumi Abayomi-Alli *

Department of Information Systems, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, LT-10223 Vilnius, Lithuania; meghalmandot97@gmail.com (M.M.)
* Correspondence: nikolaj.goranin@vilniustech.lt (N.G.); adebayo-adewumi.abayomi-alli@vilniustech.lt (A.A.A.-A.)

Abstract

The growing prevalence of online abuse has increased the need for automatic cyberbullying and hate speech detection, particularly in under-resourced and morphologically complex languages with limited annotated data and language resources. This study presents a multilingual English–Lithuanian framework combining machine learning, deep learning, transformer models, three data augmentation techniques, and explainable AI to address class imbalance, limited data, and model transparency. Overall, transformer-based approaches outperformed traditional methods. For the English dataset, averaged results over five runs with different random seeds showed model-dependent performance. Syn-Aug achieved the best F1-scores for LR (0.8791 ± 0.0029), SVM (0.8810 ± 0.0028), and RoBERTa (0.9287 ± 0.0023), while contextual BERT augmentation achieved the best F1-scores for CNN (0.9124 ± 0.0020) and BERT (0.9274 ± 0.0021). The highest English performance was obtained with the soft-voting BERT-RoBERTa ensemble, with an F1-score of 0.9356 ± 0.0035. For Lithuanian hate speech classification, hold-out validation and five-fold cross-validation were employed and evaluated on the same independent test set. For the five-fold cross-validation, XLM-RoBERTa and Syn-Aug achieved the best accuracy (0.9458) and F1-score (0.9363), followed by XLM-RoBERTa with XLM-Aug (accuracy 0.9072; F1-score 0.9336). Under the hold-out protocol, XLM-RoBERTa with XLM-Aug achieved the best accuracy (0.824) and F1-score (0.8109). These results show that Lithuanian classification performance varied with augmentation and evaluation protocol. Cross-language transfer using semantically aligned binary labels further highlighted the importance of language-specific data. Overall, the proposed approach supports efficient, interpretable, and reliable multilingual cyberbullying detection.

Keywords: cyberbullying detection; explainable artificial intelligence (XAI); hate speech detection; low-resource languages; natural language processing; transformer models

Check for updates

Academic Editors: Junaid Rashid and Patrick Siarry

Received: 17 September 2026
Revised: 30 October 2026
Accepted: 3 November 2026
Published: 4 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Social media forums have come to play an important role in the communication process among people; however, they have also made communication detrimental through practices like cyberbullying and hate speech. Such harmful contents circulate widely and victimize people, bringing about psychological trauma. Cyberbullying has emerged as an important problem within society due to its negative effects on people's psychological well-being [1].

Handcrafted identification of cyberbullying cannot be easily achieved, considering the exponential volume of social media content, especially texts; similarly, many existing automated technologies are designed to detect cyberbullying on online platforms. However, all these methods still fail to detect some of the less obvious forms of abusive behavior, making them ineffective when applied in practice [2]. Harmful content may also not be conveyed using explicitly offensive language. It may be conveyed through sarcasm, irony, coded language, and indirect attacks, which can be extremely challenging for ML models to capture. Most existing approaches to detecting online abuse have been developed primarily for the English language, limiting their effectiveness in low-resource languages. The issue of class imbalance also remains one of the biggest problems in cyberbullying datasets. Evaluating models using only a single train–test split may therefore yield unreliable results, particularly in small datasets. These problems underscore the importance of creating reliable, multilingual automated detection mechanisms to help prevent online abuse.

NLP has become an important technology that can be used to solve the issue of cyberbullying and hate speech. Despite substantial progress in cyberbullying and hate speech detection using pre-trained language models, several limitations remain in the existing literature. Most current studies are conducted in isolation of model architecture, data augmentation techniques, multilingual transfer, and interpretability. It is therefore difficult to understand how these aspects can influence each other within the same experiment. Secondly, augmentation techniques are usually tested on a relatively small number of classifiers or in a single run experiment. In addition, cross-language performance on multilingual harmful content research is often measured without considering the semantic differences in label sets, task formulation, and the properties of the languages used. Finally, explainability is often demonstrated by selected examples, with little focus on the quantification of the stability and faithfulness of such explanations.

This paper contributes a unified experimental framework for multilingual harmful-content detection as opposed to a novel classification technique by presenting (1) a controlled comparison of classical machine learning, deep learning, and transformer-based approaches for detecting cyberbullying from English texts and hate speech from Lithuanian texts; (2) a comprehensive analysis of various augmentation approaches such as synonym, generative, and contextual augmentation on different models; and (3) a cross-linguistic transfer investigation using semantically aligned binary labels, which allows us to investigate transfer performance despite the different languages, domains, and label semantics. Finally, we present an extended explainability analysis that incorporates LIME, attention, explanation stability, and controlled perturbation testing. Together, these components allow for a broader evaluation of the circumstances under which augmentation, multilingual transfer, ensembling, and explainability can be helpful in harmful content detection.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on cyberbullying and hate speech detection, while Section 3 describes the proposed multilingual framework, dataset, data augmentation strategies, and transformer-based models employed in this study. Section 4 presents the experimental results of the proposed model, explainability analysis, and comparisons with state-of-the-art methods. Section 5 concludes the paper by presenting findings and future directions.

2. Related Work

This section explores state-of-the-art techniques for the detection of cyberbullying and recent developments in deep learning models for identifying text complexity.

A study by [3] presents simple ML classifiers such as Random Forest and XGBoost for effective detection of cyberbullying on a public English cyberbullying dataset. Authors in [4] applied traditional ML models such as NB, SVM, LightGBM, and RF to detect cyberbullying

on the English Twitter dataset. Among all models tested, LightGBM exhibited the strongest performance with an accuracy of 85.5% and F1 score of 84.49%.

Convolutional Neural Networks (CNNs) have been shown to be effective at extracting local textual patterns and abusive expressions, while Bidirectional Long Short-Term Memory (BiLSTM) networks perform better in learning sequence relationships and contextual information in text [5]. Recent work from [6,7] proposed that CNN-based models perform well on explicit abusive content when trained on large datasets. BiLSTM-based models partially capture information over longer sequences [5,8]. However, despite their capabilities, recurrent neural networks are known to be computationally costly, and it is difficult for them to handle very long text sequences due to challenges like vanishing gradient problems.

To overcome some of the limitations of deep learning systems, researchers have explored hybrid systems that combine the advantages of stand-alone deep models. A hybrid architecture that integrates CNN and BiLSTM was proposed by [9,10]. A similar study by [11] implements a hybrid CNN-BiLSTM framework to detect six classes of cyberbullying behavior on the Twitter dataset written in English. Recent studies have explored the application of ensemble machine learning methods to enhance the stability and efficiency of cyberbullying detection systems. Daniel et al. [12] presented an ensemble method using tournament optimization, while [13] introduced a hybrid ensemble that includes Naïve Bayes and deep learning methods. The study showed that the proposed method achieved better classification accuracy than single models.

More recently, transformer-based architecture has gained popularity in cyberbullying and hate speech recognition because of its ability to identify complex contextual relations via self-attention mechanisms. BERT architecture was first introduced by [14]. Lately, many studies have shown that the efficiency of using transformer-based models against machine learning and deep learning algorithms has been quite rewarding. An interesting study by [15,16] demonstrated that fine-tuned BERT models outperformed conventional classifiers in recognizing explicit and implicit abusive language. Authors in [17] compared the performance of different learning models and transformer-based models for Lithuanian-language hate speech detection. State-of-the-art models such as BiLSTM, LSTM, CNN, SVM, and Random Forest models and fine-tuned Multilingual BERT, LitLatBERT, Electra, RWKV, ChatGPT (GPT-4o-mini, GPT-4, GPT-3.5-Turbo), LT-Llama2, and Gemma-2 models were employed. Subsequent studies present hybrid models based on the BERT-BiGRU-CNN model [18], and another study by [19] proposed a hybrid BERT-CNN model with multi-head self-attention mechanisms, in addition to the XLM-RoBERTa model [20], RoBERTa-BiLSTM model [17] and the DeBERTa model with a Gated Broad Learning System [21].

The rapid advancement of LLMs enabled automated text understanding and generation. Authors in Wen [22] concluded that GPT-4o achieves better performance on sensitive content detection tasks, which implies the possibilities of generative AI models for detecting cyberbullying. Another interesting study by Cirillo et al. [23] proposed an LLM-based detection against traditional ML and NLP baselines across multiple datasets, including cyberbullying detection, sarcasm recognition, sentiment analysis, and emotion classification. Their results show that LLMs can perform significantly better in these tasks. However, there are still prevailing problems with the computational costs, interpretability, and deployment efficiency of LLMs. As the models for cyberbullying and hate speech detection become more complex, so does the need to explain them. Recent work emphasizes that high accuracy is insufficient for reliable cyberbullying and hate speech detection. Thus, a model that performs well on benchmarks but cannot justify its decisions is often difficult to deploy in real moderation contexts. Authors from Nirmal et al. [24] explored rationales generated by LLMs for explainable detection of hate speech, while Calabrese et al. [25] proposed that structured explanations increase the efficiency of classification tasks. Further studies by

Salles et al. [26] illustrated that although explanations provided by LIME and SHAP could seem plausible, they were not necessarily faithful enough to the decision process of the mode. According to Hu and Lee [27], explanation quality needs to include faithfulness and causal grounding in addition to traditional classification criteria. Similarly, Cirillo et al. [23] explored prompt-based classification approaches using emerging LLMs for detecting cyberbullying in social media posts, highlighting their potential as alternatives to conventional ML-based detection methods.

There are still some prevailing challenges associated with cyberbullying detection, as identified in previous studies, such as poor representation of detection models in low-resource languages. In addition, limited datasets, the issue of class imbalance, and the absence of interpretability relate to the difficulties of interpreting model decisions. To solve some of the identified gaps, this study proposes an evaluation framework for detecting cyberbullying and hate speech in English and Lithuanian. In addition, this study incorporates XAI utilizing LIME, attention, explanation stability evaluation, and perturbation-based techniques.

3. Methodology

This section describes the proposed multilingual framework for cyberbullying and hate speech detection across English and Lithuanian, as shown in Figure 1.

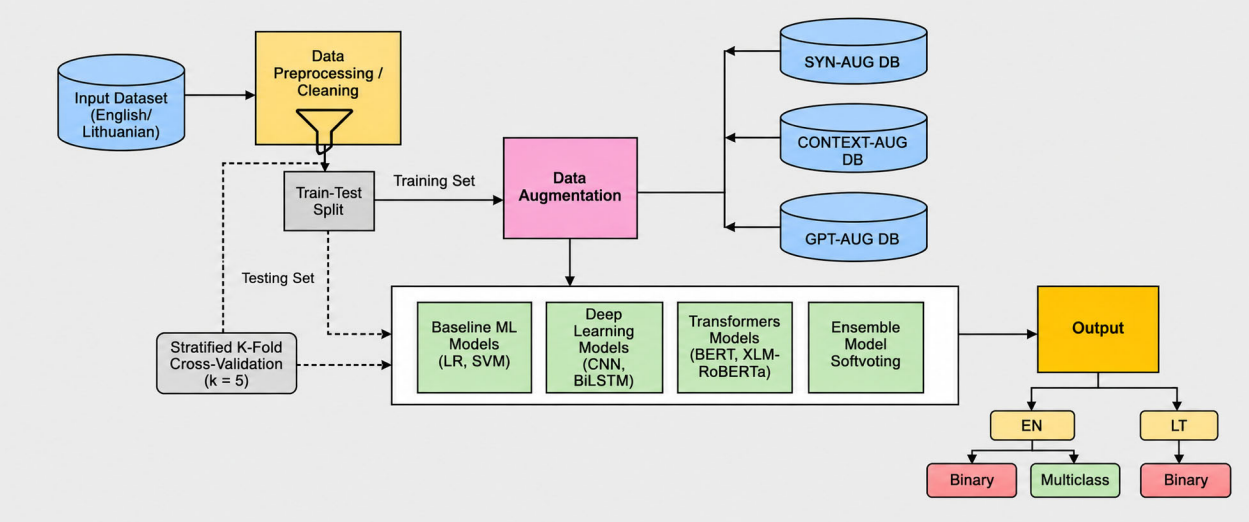


Figure 1. Proposed framework for cyberbullying detection.

3.1. Datasets and Preprocessing

The datasets used for this research are publicly available and have been chosen to provide both high-resource language (English) and low-resource language (Lithuanian) samples.

3.1.1. English Cyberbullying Classification Dataset

In this study, we used the English cyberbullying dataset proposed by [28]. The dataset contains 47,692 tweets collected from Twitter and annotated into six classes. The classes include ethnicity, age, religion, gender, other cyberbullying, or non-cyberbullying. In this study, the dataset was used for two tasks: binary and multiclass classification. We merged the five cyberbullying categories into a single class (cyberbullying, label = 1) for the binary task while keeping the non-cyberbullying class at label 0. A stratified split of 80/20 was applied; thus, 80% for the development dataset yields a development set of 38,153 records, while 20% for the independent test dataset yields a test set of 9539 records. The development

set is further divided into 32,430 training records and 5723 validation records. Notably, the independent test sets were maintained independently of any model, validation, or augmentation process. The synthetic samples were created using the training samples only.

3.1.2. Lithuanian Hate Speech Corpus V1 Dataset

This dataset was created by [29] and was collected through CLARIN-LT. The dataset contains a total of 3761 comments collected from Lithuanian online news portals and social media, labeled with a binary class (hate speech/no hate speech). The dataset shows a slight class imbalance with 1602 hate speech samples and 2159 non-hate speech samples. During data preprocessing, 14 duplicate records were removed from the non-hate-speech class, resulting in a cleaned dataset of 3747 unique instances, comprising 1602 hate speech and 2145 non-hate speech samples. The summary of the two available datasets, both in English and Lithuanian, used in this experiment is presented in Table 1, and the text preprocessing workflow is illustrated in Figure 2.

Table 1. Description of datasets used, showing class distributions and sources.

Dataset	Language	Sample	Classes	Class Balance	Source
Cyberbullying Dataset	English (EN)	47,692	Binary	Binary: Imbalanced (83.3% vs. 16.7%)	[28]
			Multiclass: 6	Multiclass: Balanced (16–17% per class)	
Lithuanian Hate Speech Corpus	Lithuanian (LT)	3761	Binary	57.4% non-hate speech; 42.6% hate speech,	[29]

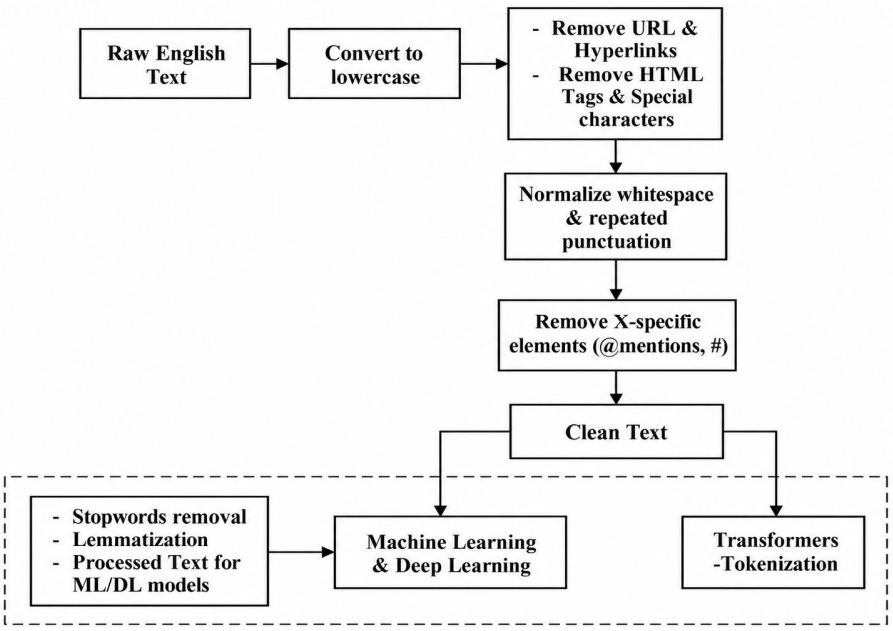


Figure 2. Pre-processing flow diagram.

To ensure an unbiased evaluation and avoid data leakage, all the dataset splits were made prior to any data augmentation process. Regarding the Lithuanian dataset, a stratified 80:20 train–test split was performed to maintain the original class distribution as depicted in the summarized Table 2. In total, 2997 samples belonged to the initial training split, while 750 samples comprised the test set. For the validation split, 15% of the original training split was taken aside through the stratified sampling method. The validation split consisted of the original samples only. Since the dataset was relatively small, data augmentation techniques

were introduced to augment the sample size and increase the variety of training samples for each class. For each augmentation approach, 1500 artificial samples were generated and added to the training set only. Thus, we had three different augmented training datasets to compare the results.

Table 2. Partitioning of the English and Lithuanian data and leakage-control procedure.

Dataset stage	**Cyberbullying (1)**	**Non-Cyberbullying (0)**	**Total**
Full English dataset	39,747	7945	47,692
Initial 80% training partition	31,797	6356	38,153
Final training partition (85% of initial training)	27,027	5403	32,430
Validation partition (15% of initial training)	4770	953	5723
Independent test partition	7950	1589	9539
# augmented samples for Syn-Aug, GPT-Aug, and contextual-Aug	-	1500 (each)	
Lithuania dataset			
Dataset stage	**Hate Speech (1)**	**Non-Hate Speech (0)**	**Total**
Full Lithuanian dataset	1602	2145	3747
Initial 80% training partition	1281	1716	2997
Final training partition (85% of initial training)	1089	1458	2547
Validation partition (15% of initial training)	192	258	450
Independent test partition	321	429	750
# augmented samples for Syn-Aug, GPT-Aug, and contextual-Aug	1500 (each)	1500 (each)	

NB: Augmentation was applied only to the training data. For English, Syn-Aug and BERT-Aug each added 1500 non-cyberbullying samples, while GPT-Aug added approximately 1000. For Lithuanian, each technique targeted 1500 samples per class (3000 total). Augmentation methods were evaluated separately, with validation and test sets unchanged.

3.2. Learning Models, Training Configuration, and Reproducibility

The experimental framework evaluated three types of learning approaches: classical machine learning classifiers, deep learning approaches, and transformer approaches. LR and SVM were selected as classical baselines, and these models used TF-IDF features as text representation and, thus, were able to learn discriminative lexical patterns without employing any learned contextual embeddings. Secondly, for the deep learning model, we employed CNN models because of their ability to learn local n-gram patterns. The third category of learning model is based on transformers, and this experiment made use of BERT and RoBERTa for the English cyberbullying detection task and XLM-RoBERTa for the Lithuanian hate speech experiment. BERT and RoBERTa offer contextualized representations for tokens that consider the linguistic context around them, whereas XLM-RoBERTa offers such representations for multilingual text and, thus, was chosen for the Lithuanian task and cross-lingual experiment. Instead of adopting one training configuration for models of different types, model-dependent representations and hyperparameters were adopted.

Each family of models had its own training settings because of their different optimization behavior and computational cost. Deep-learning models were optimized with Adam, while transformer models were fine-tuned with AdamW with lower learning rates that were suitable for pre-trained language models. An early stopping technique was used for the neural models to prevent overfitting; model selection was done based on validation performance, and the best weights were restored. For the transformer experiments, checkpoints were checked after each epoch, and the best one was selected based on the

respective validation measure. Thus, the batch size, maximum number of training epochs, learning rate, weight decay, dropout, warm-up strategy, and early stopping patience could be considered as model-specific parameters and not the same for all the experiments. The input representations, the main parameters of the architecture, optimization parameters, and training parameters used for each architecture are listed in Table 3. It is important to note that these parameters remained constant throughout the experimentation under the given conditions.

Table 3. Configuration details of the classification models used in the experiments.

Model	Configuration Parameters and Values
LR	Input representation: TF-IDF, unigram-bigram Configuration: English: max. 5000 features; Lithuanian: C = 1, L2. Max iterations = 5000 (EN), 2000 (LT); random seeds = {10, 21, 42, 77, 100}; 5 runs; solver = lbfgs (EN), saga (LT); class weight = balanced
SVM	Input representation: TF-IDF; 5000 features; unigram + bigram Configuration: LinearSVC; sigmoid calibration LinearSVC with sigmoid calibration; random seeds = {10, 21, 42, 77, 100}; 5 runs; balanced class weights; cache = 2000 MB
CNN	Input representation: Keras tokenizer; vocab = 30,000; sequence length = 120; GloVe 100-d Configuration: Conv1D = 128 filters, kernel = 3; GlobalMaxPooling; Dense = 128; Spatial dropout = 0.2; dropout = 0.3 Adam, LR = 1×10^{-4}; batch size = 128; max epochs = 80; early stopping patience = 3; random seeds = {10, 21, 42, 77, 100}; 5 runs
BERT	Input representation: bert-base-uncased; max sequence length = 128 Configuration: Sequence-classification head AdamW; LR = 2×10^{-5}; batch size = 16/64; dropout = 0.1; weight decay = 0.01; maximum epochs = 3; random seeds = {10, 21, 42, 77, 100}; 5 runs
RoBERTa	Input representation: roberta-base-uncased; max sequence length = 128 Configuration: Sequence-classification head AdamW; LR = 2×10^{-5}; batch size = 16/64; dropout = 0.1; weight decay = 0.01; maximum epochs = 3; random seeds = {10, 21, 42, 77, 100}; 5 runs
XLM-RoBERTa	Input representation: xlm-roberta-base; max sequence length = 128 Configuration: Binary sequence-classification head; maximum epochs = 5; AdamW; LR = 5×10^{-5}; batch size = 16; dropout = 0.1; weight decay = 0.01; linear scheduler; 10% warm-up; random seeds = {10, 21, 42, 77, 100};

Note: Maximum epochs represent the configured upper limit of training iterations through the dataset, rather than necessarily the number of epochs completed when early stopping is enabled.

Reproducibility and data-leakage prevention played an important role in the experimental design. The data were split into development and independent test sets, where the latter was never used for model fitting, hyperparameter tuning, and other aspects of the model selection process but was used only for assessing the final model's performance.

For five-fold cross-validation, the development dataset was divided into five stratified folds. In each iteration, the model was trained on four folds and evaluated on the remaining fold. This procedure was repeated five times, with each fold serving once as the validation set. The validation results were used to identify the optimal hyperparameter configuration for each model and augmentation strategy. Following hyperparameter selection, the final model was retrained on the complete development dataset and evaluated using the independent test set. Five test evaluations were conducted for each model configuration, with the aim of assessing the repeatability of model performance and identifying potential variations or anomalous outcomes while maintaining the same independent test set.

Furthermore, all the data augmentation steps occurred only in the corresponding training set, including within each training fold during the k-fold cross-validation. Every

model setting was evaluated using five predefined random seeds (10, 21, 42, 77, 100) to estimate the variability due to the stochastic training process, and the same independent test set was used for every seed and experiment setting. Table 3 describes the configuration details of the models used for the classification experiments, presenting the input representation/tokenizer used for each of the models, as well as the architectural details of each model. Furthermore, the table also indicates some of the main hyperparameters used in training the models, such as epochs, batch size, learning rate, hidden dimensions, optimizer, seed, early stopping metric, and dropout. In general, the settings have been chosen to keep the experiments under similar conditions for each category of models.

3.3. Data Augmentation

Augmentation was applied to the training set only to prevent data leakage, and for this study we adopted three augmentation techniques to improve the model's ability to learn underrepresented patterns.

To address the limited sample size and improve class representation, three data augmentation strategies were independently applied to the training partitions. For the Lithuanian dataset, the original training partition contained 2547 instances, comprising 1089 hate speech and 1458 non-hate speech samples. Each augmentation strategy targeted 1500 additional instances per class, yielding an expected augmented training size of 5547 samples for each technique. For the English dataset, the original training partition consisted of 32,430 instances. Synonym-based and contextual augmentation each added 1500 non-cyberbullying instances, yielding an expected training size of 33,930 samples per strategy. GPT-4o-based augmentation generated approximately 1000 additional non-cyberbullying instances, resulting in an expected training size of approximately 33,430 samples. Each augmented dataset was evaluated separately, and the validation and independent test sets were excluded from the augmentation process.

3.3.1. Synonym Replacement (Syn-Aug)

It is one of the simplest and most widely used text augmentation techniques that generates new training samples by randomly replacing selected words in a sentence. Synonym replacement was implemented in this study using the nlpaug library with WordNet. These augmenters were configured to replace at most one word per sentence with a 0.05 probability of augmentation. This implies that an approximate 5% of eligible words can be selected for replacement, thus introducing lexical variation while seeking to preserve the original semantic meaning.

3.3.2. Contextual Masked Language Model Augmentation (BERT-Aug/XLM-Aug)

BERT-Aug/XLM-Aug employs masked language models to overcome the main challenges of synonym replacement. Augmented text is generated using the semantic and syntactic contexts of a sentence rather than relying on predefined lexical resources. In this study, BERT-Aug was used for the English dataset, and XLM-Aug was applied to the Lithuanian dataset, thereby enabling context awareness.

3.3.3. GPT-Based Generation (GPT-Aug)

GPT-Aug leverages the capabilities of a generative LLM to create new training samples while maintaining the semantic intent of the original class. Unlike synonym replacement or masked language model augmentation, this strategy can generate novel, highly diverse sentences, vocabulary, and writing styles. Therefore, the use of GPT in augmentation techniques is highly useful in tackling imbalanced classes since it allows for the creation of

varied minority class instances that help to enhance the generalization of the model to new hate speech and cyberbullying.

The experimental settings for the augmentation approaches used in this study are presented in Table 4. The synonym-based augmentation uses the nlpaug to substitute words with semantically similar words, where separate configurations were used for the English and Lithuanian data sets.

Table 4. Configuration of the data augmentation strategies used in the experiments.

Augmentation Techniques	Augmented Training Size (EN/LT)	Implementation/Model	Settings
Synonym augmentation	33,930/5547	nlpaug synonym replacement	English: aug_p = 0.05, aug_max = 1; Lithuanian: aug_max = 5
BERT (Contextual) augmentation	33,930/5547	English: bert-base-uncased; Lithuanian: xlm-roberta-base embeddings	English: contextual word substitution; Lithuanian: word insertion, aug_p = 0.30, aug_max = 5
GPT-based augmentation	≈33,430/5547	GPT-4o generated synthetic text	English: approximately 1000 externally generated non-cyberbullying samples

Note: Augmentation was limited to the training data to avoid data leakage into validation and test sets.

Contextual augmentation utilized bert-base-uncased for the English language and xlm-roberta-base for the Lithuanian language through the means of contextual substitution of words and inserting words, respectively. Furthermore, the process of generating 1000 non-cyberbullying samples for the English dataset utilized GPT-4o.

To enhance the transparency and reproducibility of the GPT-based augmentation procedure, the prompt templates used with GPT-4o are provided in Appendix A. The prompts were designed to generate synthetic text consistent with the target classification labels while introducing lexical and contextual diversity. For the Lithuanian dataset, a class-conditioned prompting strategy was employed, in which the original training text and its corresponding binary label were supplied to GPT-4o to generate paraphrased samples while preserving their semantic meaning and class membership. The complete system and user prompt templates are presented in Appendix A to facilitate reproducibility. All augmentation procedures were applied exclusively to the training partition to prevent data leakage into the validation and independent test sets.

3.4. Cross-Dataset Transfer Experiment and Label Correspondence

To examine the extent to which the learned representations generalize across languages and related forms of harmful online content, a cross-dataset transfer experiment was conducted between the English Cyberbullying Classification dataset and the Lithuanian Hate Speech Corpus. The English dataset contains tweets annotated into six categories: age, ethnicity, gender, religion, other_cyberbullying, and not_cyberbullying. On GitHub, for the binary experiments, the five cyberbullying categories were merged into a single positive class (Cyberbullying), while not_cyberbullying represented the negative class. The Lithuanian corpus addresses a similar but distinct problem of hate speech detection. Hate speech is classified into six different target classes: (1) ethnicity, nationality, or race; (2) sexual orientation and gender; (3) countries; (4) political opinions; (5) religion; and (6) institutions along with neutral texts. For the CLARIN-LT Repository, the binary classification merged the six hate speech target classes into one positive class (Hate Speech) and neutral comments as a negative class (Not Hate Speech).

The relationship between the two sets of categories is highest among identity categories, and it is summarized in Table 5. The English ethnicity category has a direct conceptual link to the Lithuanian ethnicity, nationality, or race category, whereas the religion category is included explicitly in both datasets. The English gender category partly corresponds to the Lithuanian sexual orientation and gender category. Nonetheless, there are no exact pairs of categories between the other categories. The English age and other cyberbullying categories lack any Lithuanian equivalent; secondly, the Lithuanian hate speech towards countries, political views, and institutions lacks an English equivalent. Finally, the negative classes are also only operational: not_cyberbullying implies the lack of cyberbullying in the English set, whereas the Lithuanian negative class consists of neutral posts, which may even mention possible hate targets without being offensive.

Table 5. Conceptual correlation between the English and Lithuanian target labels.

English Cyberbullying Label	Lithuania Hate Speech Counterpart	Correspondence
Ethnicity	Ethnicity, nationality, or race	High
Religion	Religion	High
Gender	Sexual orientation and gender	Medium
Age	No exact match	None
Other cyberbullying	No direct correspondence.	Low
No exact match	Countries	None
No exact match	Political views	None
No exact match	Institutions	None
Not cyberbullying	Neutral/Not Hate Speech	Binary correspondence

For the cross-dataset experiment, we construct a semantically matched subset of both datasets as presented in Table 6. A binary class was constructed with target classes of positive and negative, and thus for the English dataset, ethnicity, gender, and religion make up the positive class, and not_cyberbullying depicts the negative class. The same semantic construction was done for the Lithuania dataset, where the positive class includes targets such as ethnicity/nationality/race, sexual orientation/gender, and religion. While the negative class contains neutral or not-hateful comments.

Table 6. Binary label mapping and excluded categories for semantically aligned cross-dataset transfer between English cyberbullying and Lithuanian hate speech datasets.

Basic Binary Task	English Dataset	Lithuanian Dataset
Positive	Ethnicity, Religion, Gender	Ethnicity/Nationality/Race, Religion, and Sexual orientation/Gender
Negative	Not cyberbullying	Neutral/Not Hate Speech
Excluded from experiment	Age; and other cyberbullying	Countries; Political views; Institutions

To evaluate the cross-lingual generalization capability of the proposed framework, two additional transfer experiments were conducted using XLM-RoBERTa, the only model in this study capable of processing both English and Lithuanian text. In the first experiment, XLM-RoBERTa was trained on the English training set using baseline data without augmentation and evaluated on the Lithuanian test set. In the second experiment, the model was trained on the Lithuanian training set and evaluated on the English test set. These experiments were designed to assess how well a model trained on one language transfers to the other without any target language training data, providing important context for understanding the necessity of language-specific training and augmentation strategies.

4. Results and Discussion

All the experimental findings in this section have been carried out based on the training setups tailored for each model mentioned in Section 3.2. The framework has been developed in PyTorch 2.0, Hugging Face Transformers, and TensorFlow/Keras, and all experiments were conducted on an NVIDIA Tesla T4 GPU with 16 GB GDDR6 VRAM available through Google Colab Pro. Instead of forcing one training setup on all the model families, machine learning, deep learning, and transformer-based models have been trained based on the setup appropriate for each model type. Learning rate, optimizer, batch size, max epoch number, early stopping criteria, random seed, and other replication details can be found in Section 3.2.

To ensure a consistent evaluation procedure, the same rules of data splitting were followed across all experiment configurations. Stratified sampling was applied to preserve the class distribution and avoid problems related to the class imbalance. The initial split was made with 80% for development and 20% reserved for independent testing. Validation samples were selected exclusively from the development split, while the test sample was isolated from the model training and selection processes. For the k-fold cross-validation at k = 5, the folds were created using only development data, with the independent test data remaining excluded from model training and selection. Data augmentation was also done using only the training data.

The model performance was evaluated using accuracy, precision, recall, F1 score, and ROC-AUC, while the confusion matrix was employed to assess true positives, true negatives, false positives, and false negatives. Much emphasis was attached to the F1 score due to its ability to provide a balanced assessment of precision and recall considering the class imbalance in the datasets. The experimental results for the English binary classification, English multiclass classification, Lithuanian binary classification under the evaluation protocols, and cross-dataset transfer tasks are outlined in the subsequent subsections.

4.1. Experimental Performance on English Binary Classification

This subsection demonstrates the experimental results obtained using the English cyberbullying dataset. We conducted five major experiments and compared the performances of both traditional ML (LR and SVM), deep learning, and transformer models, including CNN, BERT, RoBERTa, and BERT-RoBERTa. To take into consideration the variability in model training and to enhance reproducibility, each model–augmentation pair was trained and evaluated separately five times with random seeds 10, 21, 42, 77, and 100. For all experiments, the same fixed test set was employed. The summary of the experimental results of binary classification of the test set is provided in Table 7 and Figure 3.

Under the baseline configuration, BERT and RoBERTa generally achieved higher classification accuracy and F1-scores than the traditional machine learning and CNN models. RoBERTa achieved a slightly higher mean accuracy (0.8755 ± 0.0040) than BERT (0.8752 ± 0.0040), whereas BERT obtained a marginally higher F1-score (0.9161 ± 0.0001) than RoBERTa (0.9155 ± 0.0036). These findings indicate comparable baseline performance between the two transformer models, with only minor differences in accuracy and F1-score. The BERT-RoBERTa soft-voting ensemble achieved a numerically higher F1-score of 0.9301 ± 0.0004 and an accuracy of 0.8759 ± 0.0009, suggesting a potential benefit of combining their predictions. CNN also demonstrated competitive performance, achieving an F1-score of 0.9099 ± 0.0021, compared with 0.8738 ± 0.0029 for SVM and 0.8696 ± 0.0029 for LR. Although these results indicate a numerical improvement in F1-score, no statistical significance is inferred from the observed differences.

The Syn-Aug condition produced consistent improvements in F1-score across all five models relative to their respective baselines. The CNN model shows a smaller improvement

in the F1-score from the baseline of 0.9099 to 0.9116 ± 0.0025. The overall results obtained for the single transformer models were better, as BERT's performance improved from 0.9161 to 0.9271 ± 0.0023, which shows an improvement of ↑0.011, while RoBERTa's performance improved from 0.9155 to 0.9281 ± 0.0020. The soft-voting ensemble BERT-RoBERTa model performance increased for precision with an average score of 0.9125 ± 0.0026, compared to the baseline score of 0.9088 ± 0.0027, although the recall decreased slightly to 0.9479 ± 0.0048. In general, Syn-Aug demonstrated a good degree of variation of the lexical content, which helped achieve a better trade-off between sensitivity and control of the false-positive rate, mostly for transformer models. The GPT-Aug performance on the test dataset showed a significant drop in the performance of almost all models. CNN showed a decline in the F1-score from 0.9099 to 0.8974 ± 0.0063. A rather large variability in recall (0.8957 ± 0.0164) indicates increased sensitivity to initialization.

Table 7. Model evaluation results on the English dataset under baseline and augmented data. Reported values represent the mean ± standard deviation from five repeated runs on the same independent test set.

Baseline	Model	Accuracy	Precision	Recall	F1-Score
Baseline	LR	0.8028 ± 0.0040	0.9690 ± 0.0020	0.7886 ± 0.0045	0.8696 ± 0.0029
	SVM	0.8062 ± 0.0041	0.9552 ± 0.0022	0.8052 ± 0.0045	0.8738 ± 0.0029
	CNN	0.8484 ± 0.0034	0.9017 ± 0.0036	0.9182 ± 0.0048	0.9099 ± 0.0021
	BERT	0.8752 ± 0.004	0.8838± 0.0038	0.9509± 0.0031	0.9161± 0.0001
	RoBERTa	0.8755 ± 0.0040	0.8658 ± 0.0056	0.9713 ± 0.0030	0.9155 ± 0.0036
	BERT-RoBERTa	0.8759 ± 0.0009	0.9088 ± 0.0027	0.9524 ± 0.0025	0.9301 ± 0.0004
Syn-Aug	LR	0.8147 ± 0.0040	0.9636 ± 0.0021	0.8082 ± 0.0045	0.8791 ± 0.0029
	SVM	0.8162 ± 0.0041	0.9567 ± 0.0023	0.8164 ± 0.0044	0.8810 ± 0.0028
	CNN	0.8508 ± 0.0039	0.8999 ± 0.0040	0.9237 ± 0.0063	0.9116 ± 0.0025
	BERT	0.8759 ± 0.0039	0.9080 ± 0.0034	0.9471 ± 0.0043	0.9271 ± 0.0023
	RoBERTa	0.8767 ± 0.0036	0.9020 ± 0.0044	0.9561 ± 0.0038	0.9281 ± 0.0020
	BERT-RoBERTa	0.8761 ± 0.0018	0.9125 ± 0.0026	0.9479 ± 0.0048	0.9298 ± 0.0013
GPT-Aug	LR	0.7499 ± 0.0044	0.9813 ± 0.0017	0.7135 ± 0.0051	0.8262 ± 0.0035
	SVM	0.7615 ± 0.000	0.9626 ± 0.0022	0.7428 ± 0.0050	0.8385 ± 0.003
	CNN	0.8294 ± 0.0089	0.8993 ± 0.0055	0.8957 ± 0.0164	0.8974 ± 0.0063
	BERT	0.8509 ± 0.0055	0.9251 ± 0.0034	0.8935 ± 0.0067	0.9090 ± 0.0036
	RoBERTa	0.8628 ± 0.0039	0.9161 ± 0.0040	0.9196 ± 0.0044	0.9178 ± 0.0023
	BERT-RoBERTa	0.8572 ± 0.0034	0.9264 ± 0.0016	0.9001 ± 0.0058	0.9131 ± 0.0024
BERT-Aug	LR	0.8093 ± 0.0040	0.9642 ± 0.0021	0.8008 ± 0.0044	0.8750 ± 0.0028
	SVM	0.8111 ± 0.0040	0.9507 ± 0.0023	0.8157 ± 0.0043	0.8780 ± 0.0028
	CNN	0.8521 ± 0.0028	0.9006 ± 0.0019	0.9246 ± 0.0061	0.9124 ± 0.0020
	BERT	0.8762 ± 0.0022	0.9061 ± 0.0020	0.9499 ± 0.0045	0.9274 ± 0.0021
	RoBERTa	0.8783 ± 0.0040	0.9072 ± 0.0044	0.9513 ± 0.0051	0.9287 ± 0.0023
	BERT-RoBERTa	0.8908 ± 0.0023	0.9029 ±0.0032	0.9678± 0.0051	0.9356 ± 0.0035

Both LR and SVM were significantly impacted since there was a drop in their recall and F1 score even though the precision remained very high. One plausible reason for this degradation might be hallucination resulting in adding subtle noise, semantic drift, and task-

irrelevant variation. In addition, for this study, no formal human validation was conducted to verify the semantic consistency of the generated samples with the original samples.

The test dataset performance on BERT-Aug resulted in significant improvement for all models, which achieved higher F1-scores than their baseline counterparts. LR, SVM, and CNN models reach an F1-score up to 0.8750 ± 0.0028, 0.8780 ± 0.0028, and 0.9124 ± 0.0020, respectively. The largest performance gains were achieved by the transformer models with BERT, which demonstrated the following performance: accuracy of 0.8762 ± 0.0022, precision of 0.9061 ± 0.0020, recall of 0.9499 ± 0.0045, and F1-score of 0.9274 ± 0.0021. RoBERTa was the best-performing single model with an accuracy score of 0.8783 ± 0.0040, precision, recall, and F1-score of 0.9072 ± 0.0044, 0.9513 ± 0.0051, and 0.9287 ± 0.0023, respectively. Compared to the baseline, the improvement of the RoBERTa F1-score is about 0.0132, and the improvement of the BERT is about 0.0113.

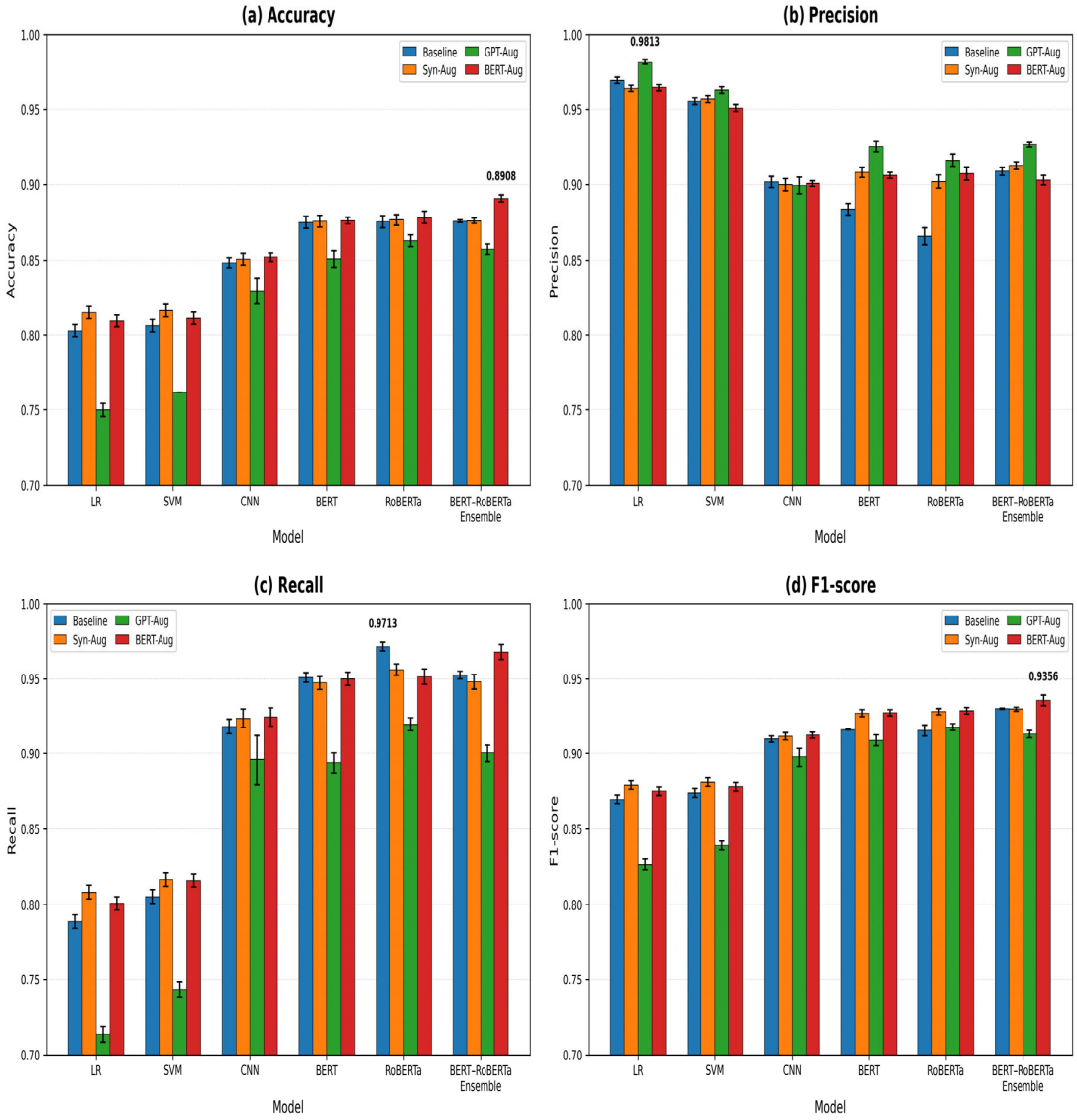


Figure 3. Comparative analysis of classification metrics for various models and augmentation approaches in English binary classification: (**a**) accuracy, (**b**) precision, (**c**) recall and (**d**) F1-score.

Figure 3 also reinforces the reported results in Table 7, with transformer-based methods achieving the top results in accuracy and F1-score compared to LR, SVM, and CNN. BERT and RoBERTa give the best and most reliable results, while Syn-Aug and BERT-Aug increase the precision-recall trade-off for transformer methods; GPT-Aug causes a definite precision-

recall trade-off. Using GPT-Aug, LR and SVM have extremely high values of precision but low values of recall, resulting in lower F1-score and lower accuracy. In general, the small error bars for most settings imply a lack of variance in five different seeds, while better F1-scores for Syn-Aug and BERT-Aug indicate more reliable augmentation methods. The high performance can be attributed to the combined strengths of BERT and RoBERTa, allowing the analysis of semantic and contextual features. The experimental results clearly show the impact of data augmentation and ensemble learning for creating an accurate and robust cyberbullying detection system.

4.2. Experimental Performance on English Multiclass Classification

The summary of the multiclass classification results is provided in Table 8. The traditional machine learning approaches, SVM and logistic regression (LR), achieved F1-scores of 0.8215 and 0.8275, respectively. A slight improvement was achieved with deep learning models, which are BiLSTM and CNN, with F1-scores of 0.8305 and 0.8419, respectively. The transformer models achieved better performance results in comparison to the traditional models, with BERT achieving an F1-score of 0.8695 and an ROC-AUC of 0.9796 and RoBERTa having a slightly better F1-score of 0.8702 and an ROC-AUC of 0.9803. In general, the best performance was achieved by the BERT-RoBERTa soft-voting ensemble with an accuracy of 0.8765, a precision of 0.8752, a recall of 0.8760, and an F1-score of 0.8750. Moreover, this ensemble had the highest ROC-AUC score of 0.9811 among all the examined models. Even though the numerical gain of the ensemble over the individual transformer-based models is insignificant, the consistent increase across all performance scores demonstrates the strengths of combining both models.

Table 8. Best performance comparison of English multiclass results.

Model	Accuracy	Recall	Precision	ROC-AUC	F1-Score
SVM	0.8225	0.8218	0.8237	0.9638	0.8215
LR	0.8278	0.8272	0.8316	0.9690	0.8275
BILSTM	0.8344	0.8340	0.8410	0.9701	0.8305
CNN	0.8440	0.8436	0.8490	0.9729	0.8419
RoBERTa	0.8719	0.8714	0.8705	0.9803	0.8702
BERT	0.8706	0.8701	0.8697	0.9796	0.8695
BERT + RoBERTa Ensemble	0.8765	0.8760	0.8752	0.9811	0.8750

For better insight into the model's behavior, we analyzed the performance of the BERT-RoBERTa soft-voting ensemble per class, as shown in Table 9. The highest performance was observed in the case of age, ethnicity, and religion-based cyberbullying, with F1 scores of 0.9862, 0.9773, and 0.9634, respectively. For gender-based cyberbullying, an F1-score of 0.9064 was obtained, while the model experienced degradation in detecting the "Other Cyberbullying" class with an F1-score of 0.7408. This could be attributed to the inherent diversity in the class, since other forms of cyberbullying involve a variety of expressions that do not easily fall under the demographic categories defined by the model. The class "Not Cyberbullying" was the most difficult class to predict with an F1-score of 0.6756. While precision was 0.7184, recall only amounted to 0.6375, which implies that quite a significant number of non-cyberbullying tweets were misclassified. Despite some of the identified differences in class-specific performance, the scores of 0.8755 for the weighted F1-score and 0.8750 for the macro F1-score indicate good performance on BERT-RoBERTa model. Hence, the similarity of the scores also suggests that no single class dominance affected overall performance. The confusion matrices of the best-performing model BERT-RoBERTa are presented in Figure 4a and 4b, respectively. From the analysis of the multiclass confusion matrix, we can see that there are more misclassified samples from Not Cyberbullying and

Other Cyberbullying classes, whereas there are fewer misclassifications with demographic categories like age, ethnicity, and religion. The ROC curve for the multiclass is depicted in Figure 5, indicating good discrimination across all classes.

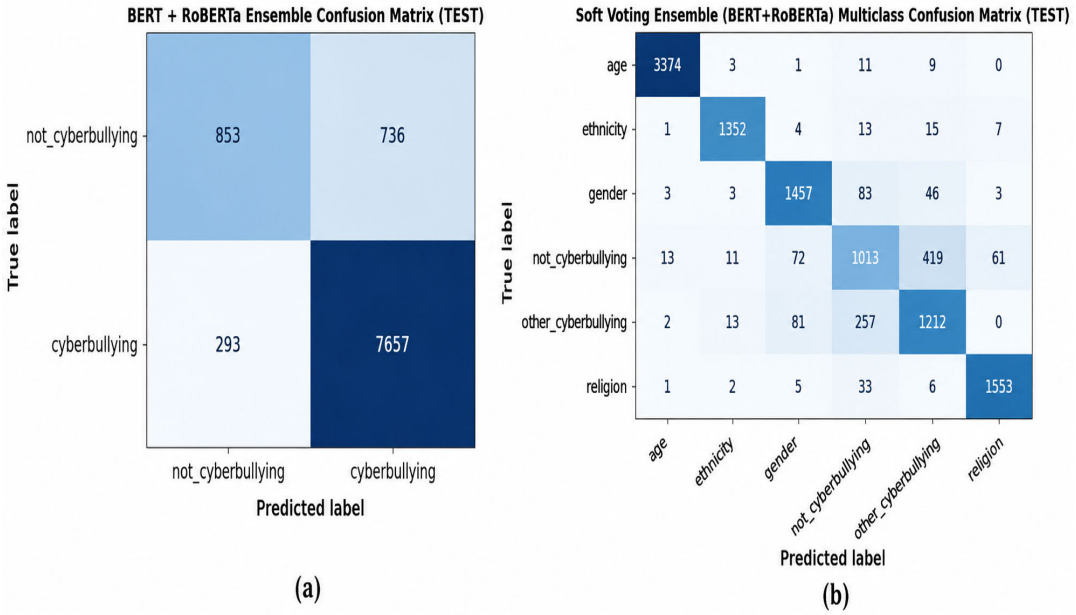


Figure 4. Confusion matrix showing class-wise performance: (**a**) binary-class; (**b**) multiclass.

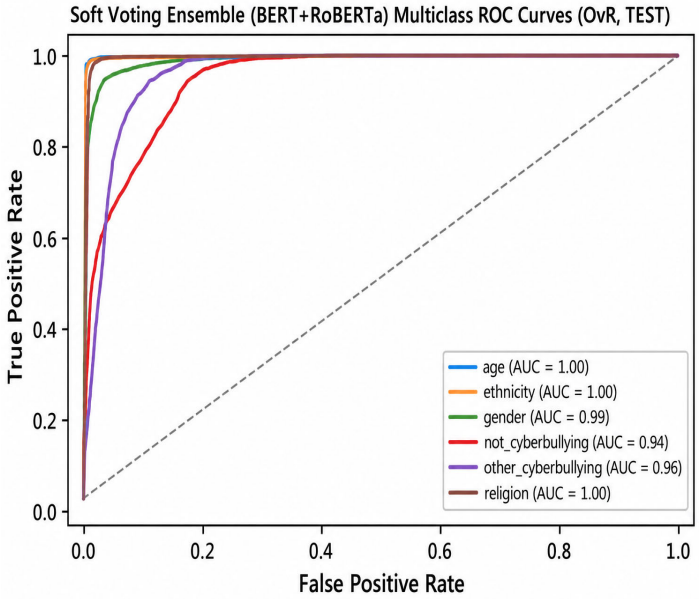


Figure 5. ROC curve of BERT + ROBERTa ensemble for English multiclass.

Table 9. Per-class performance highlighting class imbalance effects.

Class	Recall	Precision	F1-Score
Not Cyberbullying	0.6375	0.7184	0.6756
Ethnicity	0.9749	0.9798	0.9773
Age	0.9850	0.9875	0.9862
Religion	0.9706	0.9563	0.9634
Gender	0.9135	0.8994	0.9064
Other Cyberbullying	0.7744	0.7100	0.7408
Weighted Average	0.8765	0.8759	0.8755
Macro Average	0.8760	0.8752	0.8750

4.3. Experimental Performance on Lithuanian Binary Classification

For the Lithuanian dataset, Table 10 represents examples of augmented Lithuanian texts. The summary of experimental test results based on different augmentation and types of evaluation procedures is depicted in Table 11. In addition, the confusion matrix and ROC curve of the XLM-RoBERTa model for the Lithuanian binary classification test set are shown in Figure 6.

Table 10. Representative samples from Lithuanian to demonstrate the outputs of Syn-Aug, XLM-Aug, and GPT-Aug in cases of non-hate-speech and hate speech categories.

Original Class	Original Text	Syn_Aug	XLM_AUG	GPT_Aug
Non-hate speech (0)	švedija šiais laikais tampa alternatyva šveicarijai šiaur?s regione pasidom?kite ekonomikos naujienomis bepigu jiems vieni kitus provokuoti	švedija šiais laikais tampa pasirinkimas šveicarijai šiaur? s regione pasidom? kite ekonomikos naujienomis bepigu jiems vieni kitus provokuoti	švedija šiais vasaros laikais vis tampaalternatyva šveicarijai šiaur?s regione tad pasidom?kite ekonomikos naujienomis bepigu leis jiems vieni kitus provokuoti	Šiais laikais Švedija Šiaurės regione vis dažniau vertinama kaip alternatyva Šveicarijai. Pasidomėkite ekonomikos naujienomis-jiems nesunku vieniems kitus provokuoti.
Hate speech (1)	kai dekanozov ikiso paleckiui pajalnika i subine kad asmetona surakintas grandinemis nebuvo pristatytas i kremliu tada ir atsirado istorija apie upeli kacapu propaganda daugiau nieko	kai dekanozov ikiso paleckiui pajalnika i subine kad asmetona surakintas grandinemis nebuvo pristatytas i kremliu tada ir atsirado pasakojimas apie upeli kacapu propaganda daugiau nieko	tik kai mano dekanozoviui ikiso paleckiui pajalnikai subinedavo kad asmetona surakintas grandinemisnebuvo pristatytas i kremlius tada ir atsiradoistorija apie upeli kacapu propaganda daugiau nieko	dekanozovas paleckiui parode savo vieta nes smetona surakintas taip ir nebuvo nutemptas i kremliu o ta pasaka apie upeli atsirado veliau, eilinė kacapu propaganda ir daugiau nieko

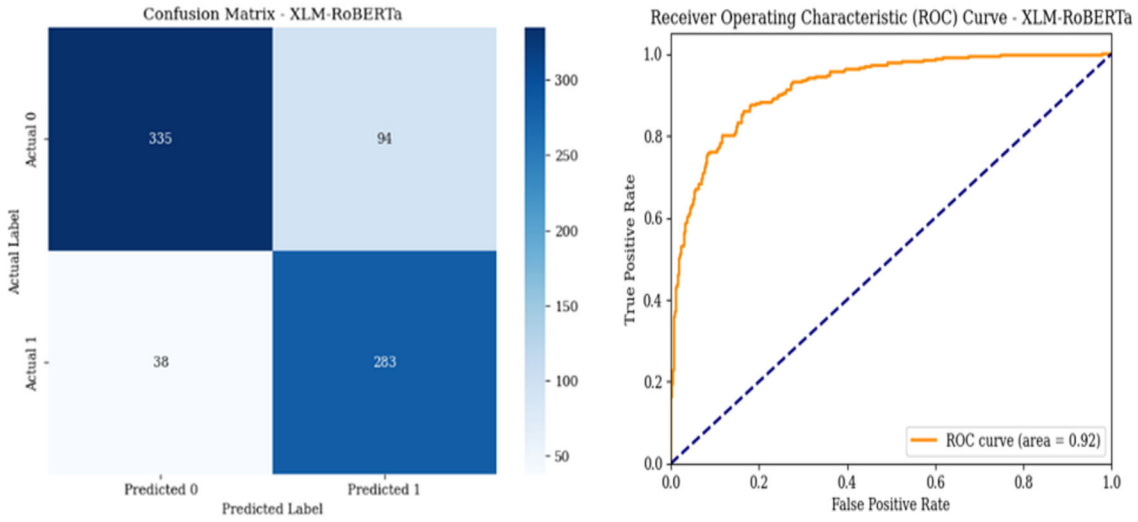
Table 11. Best observed performance of Lithuanian binary classification models on the independent test set across five test evaluations, following hold-out and five-fold cross-validation model selection.

Augmentation	Model	Independent Test (Hold-Out)		Independent Test (5-Fold CV Protocol)	
		Accuracy	F1-Score	Accuracy	F1-Score
Baseline	LR	0.7453	0.6508	0.7204	0.6718
	SVM	0.7547	0.6871	0.7234	0.6557
	CNN	0.7427	0.7174	0.5742	0.0201
	XLM-RoBERTa	0.7907	0.7573	0.8298	0.7973
Syn-Aug	LR	0.7373	0.6936	0.8427	0.8165
	SVM	0.7307	0.6814	0.9166	0.9019
	CNN	0.7547	0.727	0.6336	0.1686
	XLM-RoBERTa	0.572	0	0.9458	0.9363
GPT-Aug	LR	0.732	0.6864	0.7275	0.6875
	SVM	0.7373	0.6917	0.8044	0.7658
	CNN	0.768	0.7138	0.6247	0.1576
	XLM-RoBERTa	0.808	0.7825	0.5782	0.0859
XLM-Aug	LR	0.7253	0.6879	0.8136	0.8642
	SVM	0.7307	0.6844	0.8707	0.9092
	CNN	0.7693	0.6802	0.7113	0.8269
	XLM-RoBERTa	0.824	0.8109	0.9072	0.9336

Table 11 shows the summary of the best observed experimental results on the independent test datasets for hold-out and 5-fold CV evaluation. In the holdout, XLM-RoBERTa consistently gave better performance. F1-score is calculated using binary averaging, with hate speech (label 1) as the positive class. As compared to baseline XLM-RoBERTa, with an accuracy and F1-score of 0.7907 and 0.7573, respectively, the addition of GPT-based augmentation gave the following scores: accuracy of 0.808 and F1-score of 0.7825. The

highest performance with respect to hold-out evaluation is given by the combination of XLM-RoBERTa with contextual augmentation (XLM-Aug), achieving an accuracy of 0.8240 and F1-score of 0.8109.

The results shows substantial performance differences between the two evaluation protocols. The baseline CNN achieved 0.5742 accuracy but only 0.0201 F1 under five-fold evaluation, suggesting poor recognition of hate speech. XLM-RoBERTa with Syn-Aug obtained 0.5720 accuracy and 0.0000 F1 in hold-out testing. Since non-hate-speech samples constitute 57.2% of the test set, these values are consistent with predicting all instances as non-hate speech. These discrepancies warrant examining class-specific predictions and evaluation procedures before attributing performance differences to augmentation or assuming greater stability under cross-validation.



(**a**) (**b**)

Figure 6. Performance evaluation of the XLM-RoBERTa model on the Lithuanian binary classification test: (**a**) confusion matrix; (**b**) ROC-AUC curve.

However, a different trend was observed in k-fold cross-validation, where XLM-RoBERTa with synonym augmentation (Syn-Aug) showed the best results in terms of accuracy (0.9458) and F1-score (0.9363). In addition, the contextual augmentation was also efficient, as the second-best score was achieved by XLM-RoBERTa with contextual augmentation, which had an accuracy of 0.9072 and an F1-score of 0.9336. Synonym augmentation helped SVM to reach an accuracy of 0.9166 and an F1-score of 0.9019, while synonym-based and contextual augmentation can be extremely helpful for Lithuanian classification.

In contrast, GPT augmentation resulted in substantially lower cross-validation performance of XLM-RoBERTa with an accuracy of 0.5782 and an F1-score of 0.0859 even though it performed effectively in hold-out testing. In general, the results indicate that the augmentation effectiveness is dependent on the type of classifier used and the test method. However, synonym and contextual augmentation proved to be the most effective approaches in the Lithuanian dataset, with synonym being the most effective approach in cross-validation and contextual being the most effective approach in hold-out testing.

4.4. Performance Results on Cross-Lingual

The cross-lingual experiments show a notable asymmetry in the zero-shot transfer from English to Lithuanian. To minimize task disparity, the dataset is converted into a binary label representation (positive and negative classes) as described in Section 3.4. XLM-RoBERTa was used for both transfer methods to keep the multilingual nature of the architecture

unchanged. The summary of the zero-shot transfer is presented in Table 12, where the accuracy and F1-score for the transfer from English to Lithuanian were 0.5707 and 0.3374, and vice versa for Lithuania to English, achieving an accuracy of 0.5038 and an F1-score of 0.5878.

Table 12. Zero-shot cross-lingual transfer performance of XLM-RoBERTa.

Train Dataset	Test Dataset	Model	Accuracy	F1-Score
English	Lithuanian	XLM-Roberta	0.5707	0.3374
Lithuanian	English	XLM-Roberta	0.5038	0.5878

Figure 7 further demonstrates the scale of the transfer gap compared to performance on domain. In-domain English and Lithuanian experiments a significant F1 score was achieved for English and Lithuania with values of 0.9356 and 0.9363, respectively. However, in both cases of zero-shot transfer, there is a considerable drop in performance with a larger degradation in the case of English-to-Lithuanian transfer which reveals that representations acquired using the English corpus fail to transfer to Lithuanian without adaptation to the target language. The obtained results cannot account merely for the discrepancy in label space, as have excluded non-corresponding categories before transfer, or for differences in architecture, which is not present in the experiment.

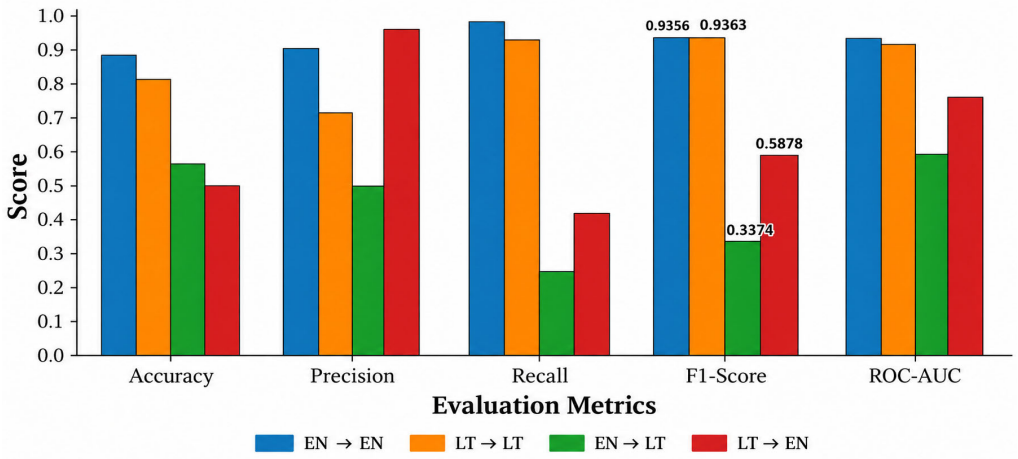


Figure 7. In-domain and zero-shot cross-lingual performance for the semantically aligned English and Lithuanian binary classification tasks.

The experimental results are analyzed in the context of domain and distributional shift. While the target classes were matched, it should be noted that the English and Lithuanian datasets originate from different online platforms and thus have different lexicon usage, discourse markers, and culturally distinct ways of expressing hate or abusive behavior. This means that zero-shot learning entails generalization from one language, vocabulary, morphological structure, and domain to another. Moreover, the fact that there is a directional asymmetry of English-to-Lithuanian and Lithuanian-to-English transfer indicates that the problem of transfer is not symmetrical and may depend on representation coverage of the target domain.

However, the model performed slightly better with a good F1-score of 0.5878 and a significant precision score of 0.9555 when trained on Lithuanian and tested on English (LT → EN). Nevertheless, the model had a considerably poor recall score of 0.4244, which implies that more than half of the cyberbullying cases were wrongly classified. Some conclusions from the cross-lingual evaluation show that multilingual transformer-based models cannot effectively handle certain linguistic and morphological differences across

languages. In addition, the efficiency of the transfer learning is determined not only by the type of architecture but also by the amount of data, similarities of languages, and the inclusion of cyberbullying expressions. Finally, the results show that cross-lingual methods alone are insufficient to build robust detectors without language-specific annotated data. Large difference in results of monolingual and cross-lingual experiments also emphasizes the need to collect and annotate specific data for poorly studied languages like Lithuanian.

4.5. Comparison with State-of-the-Art Methods

To evaluate the robustness of our proposed approach, we compared it with the recent cyberbullying and hate speech detection studies that utilized the same English dataset. As shown in Table 13, the previous methods include traditional machine learning approaches, feature-engineering techniques, data augmentation methods, deep learning models, and transformer-based architectures. The BERT-RoBERTa model performed comparably well to existing literature in the English cyberbullying dataset and, in some instances, even better. Nonetheless, these comparisons are interpreted with caution since previous work could have used different methods regarding data splitting, preprocessing, augmentation, hyperparameters, and evaluation process. Therefore, this comparison indicates strong relative performance on the shared benchmark.

Table 13. Comparison with state-of-the-art cyberbullying and hate speech detection methods on the benchmark English dataset.

Authors	Method	Accuracy	Recall	Precision	F1-Score
[30]	SparkNLP + nlpaug	0.890	-	-	-
[31]	CNN/LSTM/BiLSTM/BiLSTM-Attention	0.940	-	-	-
[3]	RF, XGBoost	0.930	0.930	0.930	0.930
[32]	TF-IDF + SVM	0.920	0.920	0.920	0.920
[33]	GAN-Augmented BERT (GAB)	0.593	-	-	0.516
[34]	SCV/SVM + TF-IDF	0.798	0.798	0.826	0.796
Our Work	BERT + RoBERTa Ensemble	0.8908± 0.0023	0.9678 ± 0.0051	0.9029 ± 0.0032	0.9356 ± 0.0035

The proposed BERT-RoBERTa soft-voting ensemble achieved a mean F1-score of 0.9356 ± 0.0035 across five independent runs using contextual BERT-based augmentation, demonstrating competitive performance against existing approaches [3,30] that reported F1-scores of 0.93 and 0.92, respectively, on the English cyberbullying dataset. Although [31] reported a higher accuracy of 0.94 using a combination of BiLSTM and BERT models, our approach still achieved a stable, higher recall of 0.9678 ± 0.0051 while maintaining a precision of 0.9029 ± 0.0032 and an F1-score of 0.9356 ± 0.0035. Although our method achieves a slight performance improvement, there is a challenge in identifying implicit forms of cyberbullying and differentiating between victims' narratives from abusive content, thus showcasing the need for context-aware solutions.

4.6. Explainability Analysis and Interpretation

This subsection presents the analysis of the explainability of the proposed BERT-RoBERTa soft-voting ensemble for English cyberbullying detection. For a better understanding of the decision-making mechanism of the model, the technique of explainable artificial intelligence (XAI) has been utilized, which includes attention weight visualization and Local Interpretable Model-Agnostic Explanations (LIME).

4.6.1. LIME Analysis of Correct and Incorrect Predictions

To complement the overall performance metrics and for a better understanding of the local reasoning process that the model employs during prediction, the Local Interpretable Model-agnostic Explanations (LIME) approach was explored, using both correct and incorrect predictions made by the BERT-RoBERTa soft-voting ensemble as depicted in Figure 8a–d. The examination of LIME involved taking into consideration all four possible outcomes of the confusion matrix, namely the true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs), to identify what specific lexical word contributed to the respective outcomes. Figure 8a–d presents samples of XAI for decision support showing both correct and incorrect predictions. Examining both correct (TP and TN) and incorrect (FP and FN) predictions is important for assessing the effectiveness of automated detection systems, especially in high-stakes real-life environments. LIME was used to estimate the contribution of individual words to each prediction by assigning high importance scores to words and expressions with strong semantic connections to abusive, offensive, and identity-based content. The explainability analysis and attention-weight visualization in Figure 8 show that the model relies primarily on semantically relevant contextual information and offensive phrases.

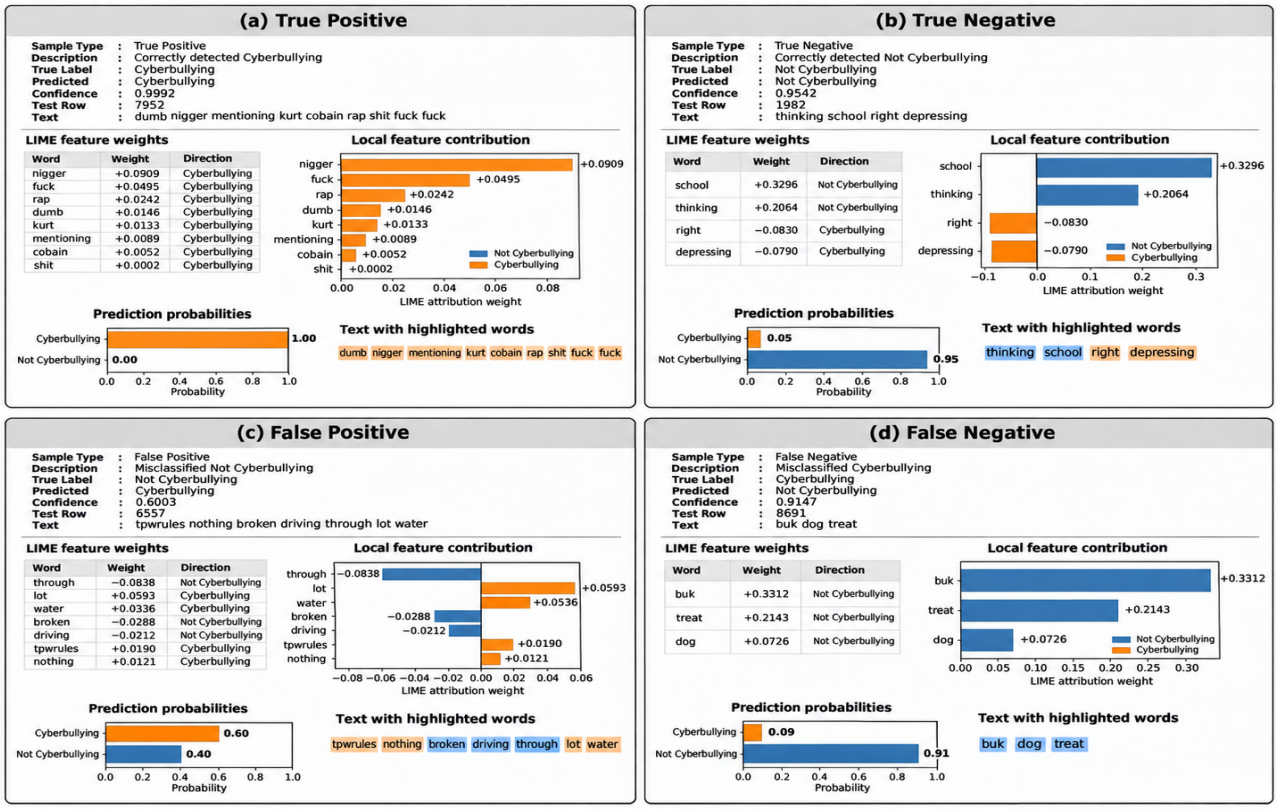


Figure 8. LIME explanations for best model predictions across the four confusion matrix results: (**a**) true positives (TPs), (**b**) true negatives (TNs), (**c**) false positives (FPs), and (**d**) false negatives (FNs).

The explainability analysis, as depicted in Figure 8a, highlights the TP instance, showing that the prediction is made based on only a few heavily weighted abusive words, along with other offensive words, adding to the same effect. The probability of cyberbullying can be attributed to the LIME explanation, which justifies that decision of using clear and heavy-weighted language. Similarly, examples of true negatives are presented in Figure 8b, showing the words that contribute significantly to the classification as "Not Cyberbullying", while there are relatively few words that give weak contrary evidence to cyberbullying.

The misclassification examples as depicted in Figure 8c,d for FPs and FNs demonstrate limitations of the model, where the FP scenario shows how the model wrongly predicts the not_cyberbullying to cyberbullying class. The LIME explanation displays some conflicting contributions of both positive and negative with some rather normal lexical features that receive significant positive weight that influences the decision as a harmful class. The FN case shows how harmful content can go unnoticed by the classifier in cases where the non-harmful or context-independent properties outweigh the locally represented content. The highest LIME contribution in this case favors the not_cyberbullying class, while the terms responsible for cyberbullying are relatively low. These examples are major constraints in existing transformer-based cyberbullying detection models. The models can detect lexical cues characteristic of abusive language, but it might be challenging to understand the overall context and intentions behind the message. In this case, the use of cyberbullying terms caused the model to classify the message as an act of cyberbullying, even though the author did not actually bully anyone; he just described the results of bullying. In addition, attention weight visualization is presented in Figure 9, which shows how the BERT-RoBERTa soft-voting ensemble gives attention to each token while predicting binary cyberbullying.

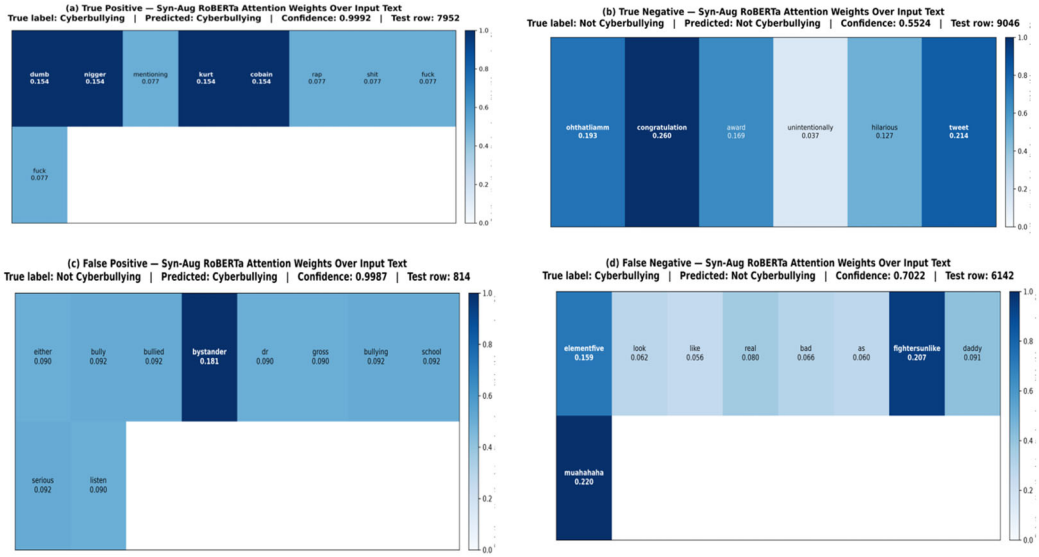


Figure 9. Visualization of attention heatmap weights over input English text—final layer CLS token averaged across all attention heads.

The attention maps in Figure 9 demonstrate the presence of token-level attention differences between Syn-Aug RoBERTa's correct and incorrect predictions. For the true positive case, the token-level attention is mainly on the abusive vocabulary, corresponding to the correct prediction of cyberbullying in the sentence, while for the true negative case, attention is paid to a few neutral tokens, which corresponds to the correct non-cyberbullying prediction. In addition, for the false positive case, the attention is on a few contextually benign tokens, which implies that the model tends to over-focus on some words. Similarly, for the false negative case, the attention is on the wrong tokens but not on potentially harmful context. Thus, these attention patterns show how misclassification may occur due to wrong representation focus.

Local explanations generated by the Lithuanian XLM-RoBERTa classifier are presented in Figure 10a–d, illustrating examples of the four possible results: true positive, true negative, false positive, and false negative. In Figure 10a, the model predicted the text to be hate speech with an extremely high level of confidence. It is evident that a few highly weighted words pushed the prediction towards the hate speech class, whereas other words contributed only modestly. Example Figure 10b is a case of TN correctly classified as not hate speech

class; hence, the significant word in this sample carries minor contribution, and the sentence is more like a conversation and not a threat.

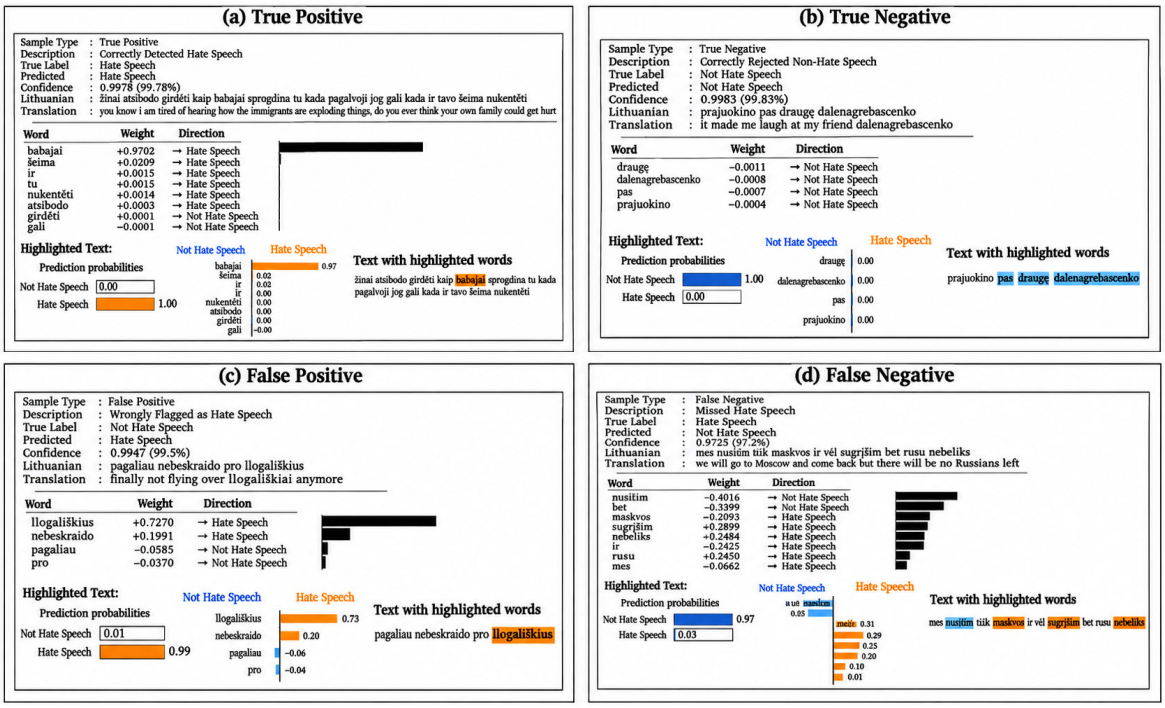


Figure 10. Explanation of local representative instances of predictions by the XLM-RoBERTa model for Lithuanian hate speech classification. The four subplots are an example of (**a**) true positive, (**b**) true negative, (**c**) false positive, and (**d**) false negative.

Therefore, it seems that the model can distinguish normal references to people or places from abusive or harmful content in the absence of aggression or group-directed threats. An example of a false positive is presented in Figure 10c where "not hate speech" is misclassified as hate speech with extremely high confidence. A similar instance is depicted in Figure 10d, where a hate speech instance containing hostile language is predicted as not hate speech with an assigned high confidence score. In the former, it is clear from the local interpretation that the model is making the prediction based on the rare/local features "liogailiskius" and "nebeskraido" and not the features "pagaliau" and "pro" that support the other class. While for the latter, it seems that various individual tokens work towards the negative class label. Thus, the harmful meaning emerges from the combined interpretation of the entire utterance, rather than from one overtly abusive keyword. The attention heatmap weights visualized on the input Lithuanian text—the final layer CLS tokens averaged over all attention heads are presented in Figure 11.

In terms of practical implications, these results further highlight the importance of incorporating explainable AI algorithms in content moderation processes. The ability to observe how tokens contribute to the result provides useful insight into how the model makes its decisions, thus enabling identification of the reasons behind any errors and supporting adjustment to techniques in social media content detection. The combination of high-performing transformer models with human oversight will provide better contextual understanding that current machine learning models still lack and thus will ensure a minimal misclassification rate. Importantly, these local explanations are indications of the features associated with individual model predictions rather than causal explanations of the classifier's decision process. These attention weights should be interpreted as indicators of where the models allocate representational focus, rather than as direct causal explanations of the predictions.

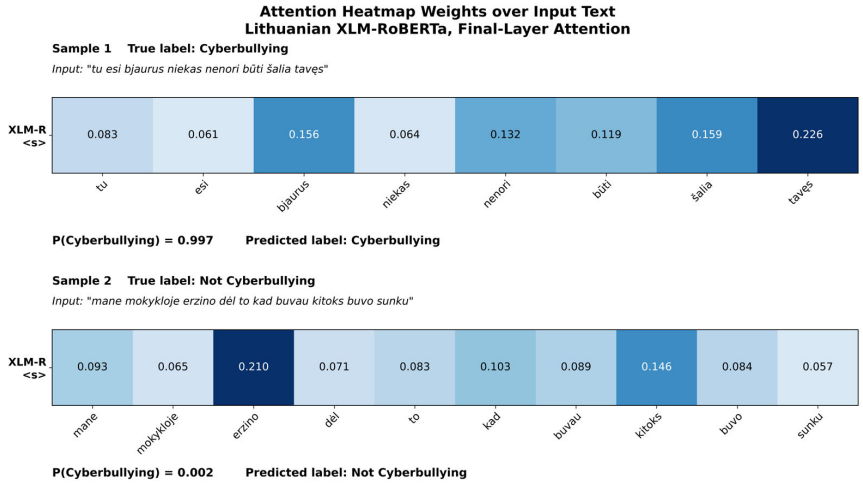


Figure 11. Visualization of attention heatmap weights over input Lithuanian text—final layer CLS token averaged across all attention heads.

4.6.2. Explanation Stability, Reliability, and Outcome Consistency

To assess the reliability of LIME explanations across different prediction outcomes, explanation stability was evaluated separately for true positive (TP), true negative (TN), false positive (FP), and false negative (FN) instances. This analysis is interpreted as consistency within outcome groups, and the goal is to determine whether explanations acted similarly both in the case of correct and incorrect prediction outcomes. Following the perturbation-based faithfulness measure for explainability in NLP [35], the effect of the LIME-identified features was examined through measuring the change in model confidence after progressively removing the highest-ranked features supporting the original prediction. Given an input x, the original prediction is denoted as $\hat{y}$. The original confidence can be expressed as $C_0 = P_\theta(\hat{y}|x)$. After cumulative elimination of the top$-k$ LIME-ranked supportive features, the confidence is recalculated as $C_k = P_\theta(\hat{y}|x^{(-k)})$, where $x^{(-k)}$ denotes the perturbed instance. The resulting confidence values C_0, C_k drops, presents a confidence change $\Delta C_k = C_0 - C_k$, indicating that the eliminated features had a stronger influence on the original prediction. This procedure is conceptually consistent with the metric of comprehensiveness introduced [35], which evaluates and measure the faithfulness after removal of explanatory features. Prediction stability T_k in Equation (1) is assigned the value of 100% when the new predicted class $\underset{y}{arg\ max} P_\theta(y|x^{(-k)})$ after the perturbed input retains the original predictions and T_k becomes 0% when the predicted class changes after perturbation differs from the original predicted class.

$$T_k = \begin{cases} 100\%, & if\ \underset{y}{arg\ max} P_\theta(y|x^{(-k)}) = \hat{y}, \\ 0\%, & Otherwise \end{cases} \quad (1)$$

The results on Table 14 reveal varying levels of prediction stability across the TP and TN instances. For rows 491 (TP) and 2482 (TN), the original predicted classes remained unchanged after cumulative removal of the top three supporting features, although confidence decreased from 0.999328 to 0.832172 and from 0.957474 to 0.842479, respectively. In contrast, the TP instance in row 8311 retained its original prediction after removing the top two features but changed class following removal of the third feature, with confidence decreasing from 0.999328 to 0.260610. Similarly, the TN instance in row 7224 changed class after removing the top two features, as confidence declined from 0.922909 to 0.454785, and it remained changed after the third removal (0.378681). These findings demonstrate that the influence of LIME-identified features varies across individual instances, with some

predictions remaining robust while others are sensitive to the removal of a small number of influential features.

Table 14. Outcome-stratified perturbation-based faithfulness analysis of LIME explanations for the Syn-Aug RoBERTa model under cumulative removal of the top-ranked class-supporting features.

Row ID	Outcome	Predicted Class	Original Confidence	Perturbation Level (Remove)	Removed Words and Lime_Weights	Confidence After Removal	Prediction Stability (%)
491	TP	Cyberbullying	0.999328	Remove Top-1	racist (0.0516)	0.999252	100%
	TP		0.999328	Remove Top-2	racist (0.0516), racism (0.0476)	0.986495	100%
	TP		0.999328	Remove Top-3	racist (0.0516), racism (0.0476), colored (0.0406)	0.832172	100%
8311	TP	Cyberbullying	0.999328	Remove Top-1	gay (0.1130)	0.999239	100%
	TP		0.999328	Remove Top-2	gay (0.1130), rape (0.1002)	0.828851	100%
	TP		0.999328	Remove Top-3	gay (0.1130), rape (0.1002), joke (0.0722)	0.260610	0%
2482	TN	Not Cyberbullying	0.957474	Remove Top-1	college (0.1996),	0.912277	100%
	TN		0.957474	Remove Top-2	college (0.1996), theatre (0.0697)	0.881095	100%
	TN		0.957474	Remove Top-3	college (0.1996), theatre (0.0676), year (0.0425)	0.842479	100%
7224	TN	Not Cyberbullying	0.922909	Remove Top-1	school (0.2057)	0.892606	100%
	TN		0.922909	Remove Top-2	school (0.2057), thanksgiving (0.1484)	0.454785	0%
	TN		0.922909	Remove Top-3	school (0.2057), thanksgiving (0.1484), lucky (0.0978)	0.378681	0%
788	FP	Cyberbullying	0.957936	Remove Top-1	fucking (0.2679)	0.840002	100%
	FP		0.957936	Remove Top-2	fucking (0.2679), black (0.1936),	0.132872	0%
	FP		0.957936	Remove Top-3	fucking (0.2679, black (0.1936), using (0.0390)	0.095656	0%
814	FP	Cyberbullying	0.998681	Remove Top-1	bully (0.4027)	0.102867	0%
	FP		0.998681	Remove Top-2	bully (0.4027), school (0.1195)	0.115263	0%
	FP		0.998681	Remove Top-3	bully (0.4027), school (0.1195), bullied (0.0943)	0.158675	0%
617	FN	Not Cyberbullying	0.870597	Remove Top-1	isi (0.4299)	0.424989	0%
	FN		0.870597	Remove Top-2	isi (0.4299), take (0.0604)	0.363758	0%
	FN		0.870597	Remove Top-3	isi (0.4299), take (0.0604), nice (0.0368)	0.338456	0%
1071	FN	Not Cyberbullying	0.803961	Remove Top-1	comp (0.1059)	0.725738	100%
	FN		0.803961	Remove Top-2	comp (0.1059), lucaswj (0.0920)	0.722581	100%
	FN		0.803961	Remove Top-3	comp (0.1059), lucaswj (0.0920), year (0.0651)	0.779151	100%

For the row_id 788, an instance of an FP, removal of the highest-ranked feature brought down the confidence to 0.840002, compared with an original confidence of 0.957936, without affecting the prediction; elimination of the top two features resulted in a class change, which continued when the third-ranked feature was removed. Thus, this implies that misclassification of the cyberbullying class depends on a small number of features whose weights are very high. Finally, for row_617, an instance of an FN, the removal of the single highest-ranked feature of LIME resulted in a change in prediction from Not Cyberbullying, and the new prediction persisted despite further perturbation. This behavior indicates that the misclassified instance was supported by a narrow and unstable set of features despite its relatively high initial confidence. In conclusion, prediction stability and explanation fidelity are complementary rather than equivalent properties. While predictions can still be class stable despite a significant drop in confidence for correct predictions, incorrect predictions can change when a relatively few numbers of influential features are removed. The results therefore provide qualitative evidence that LIME is identifying features with genuine local influence on model behavior.

Limitations of this study include an insufficient dataset for the LT language, thus resulting in poor transfer, which reflects language mismatch and dataset/task mismatch rather than purely linguistic differences. Secondly, the morphological complexity and tokenizer behavior were not isolated experimentally; hence, no direct measurements were taken for subword segmentation and morphology-dependent representation quality. A further limitation is that the cross-lingual analysis was based primarily on aggregate performance metrics. Future work should incorporate class-level metrics, confusion-matrix analysis, and threshold calibration to better characterize directional transfer errors.

5. Conclusions

This study addresses some of the prevailing challenges in cyberbullying and hate speech detection, particularly the limited availability of effective solutions for low-resource and morphologically rich languages, class imbalance, and the lack of transparency in automated moderation systems. To address some of these shortcomings, we proposed a multilingual cyberbullying detection framework using both English and Lithuanian datasets. The proposed approach combined three distinct data augmentation techniques with state-of-the-art methods and transformer-based models. To further interpret the overall performance, an explainable AI technique was introduced to enhance model interpretability and support decision-making.

The results showed that transformer architecture consistently outperformed traditional machine learning and deep learning methods, especially the BERT-RoBERTa soft-voting ensemble, which achieved the best F1-score of 0.9356 ± 0.0035 for the English dataset. The same BERT-RoBERTa ensemble model also performed best on the English multiclass classification task with an accuracy of 0.8765 and an F1-score of 0.8750. The averaged experimental result for the English dataset from five runs on different random seeds proved to be model-dependent and shows the best F1-score obtained using Syn-Aug for LR (0.8791 ± 0.0029), SVM (0.8810 ± 0.0028), and RoBERTa (0.9287 ± 0.0023), while contextual-based (BERT) augmentation achieved the best F1-score with CNN (0.9124 ± 0.0020) and BERT (0.9274 ± 0.0021).

For the Lithuanian dataset, hold-out validation and five-fold cross-validation were employed as alternative model-selection protocols, with the resulting models evaluated on the same independent test set. The experimental result on the independent test set following the hold-out validation protocol showed that XLM-RoBERTa and XLM-based augmentation achieved an accuracy of 0.8240 and F1 score of 0.8109. Meanwhile, the independent test set evaluation on five-fold cross-validation protocol proved to be more effective: XLM-RoBERTa

with the synonym-based augmentation achieved the best accuracy score of 0.9458 and F1 score of 0.9363, followed by XLM-based augmentation with an accuracy of 0.9072 and F1 score of 0.9336. These values represent the best observed independent-test performance across five test evaluations. These contrasting results suggest that conclusions regarding augmentation effectiveness should be supported by multiple evaluation protocols rather than relying on a single experimental configuration, especially for low-resourced languages.

Furthermore, the cross-dataset experiment was performed on reconstructed English and Lithuanian datasets with semantically matching positive and negative target classes. This result shows that both lexicon-based and contextual augmentation can be beneficial to the task of classification of offensive content, but the overall success depends on the language features of the used dataset. The cross-transfer performance highlights the combined effects of language and domain, along with cross-dataset distributional differences, without the need to attribute performance loss simply to linguistic and morphological dissimilarities.

Finally, this study employed LIME-based explanations to support the interpretation of model predictions and identify influential lexical and contextual features. These explanations provide insights into individual predictions, although they do not independently establish model reliability or guarantee faithful representations of the underlying decision process. Overall, the findings demonstrate the potential of combining multilingual transformer models, data augmentation, and explainable AI techniques to support more transparent cyberbullying and hate speech detection. Future research will incorporate independent human evaluation of GPT-generated samples to assess label preservation, semantic drift, and inter-annotator agreement. Additional investigations will explore few-shot adaptation and machine translation for cross-lingual transfer, considering potential translation errors and semantic distortion. Finally, domain-adaptive pretraining of XLM-RoBERTa using unlabeled Lithuanian social-media text will be investigated to improve representation quality before supervised fine-tuning.

Author Contributions: The individual contributions of the authors to the conceptualization, methodology, software development, validation, investigation, manuscript preparation, supervision, and project administration are detailed as follows: Conceptualization, M.M. and O.O.A.-A.; methodology, M.M., O.O.A.-A., N.G. and A.A.A.-A.; software, M.M. and O.O.A.-A.; validation, M.M. and O.O.A.-A.; investigation, M.M. and O.O.A.-A.; writing—original draft preparation, M.M. and O.O.A.-A.; writing—review and editing, M.M., O.O.A.-A., N.G. and A.A.A.-A.; supervision, O.O.A.-A.; project administration, O.O.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original datasets analyzed in this study are publicly available. The English Cyberbullying Classification dataset is available on Kaggle at Cyberbullying Classification. The Lithuanian Hate Speech Corpus v.1 is openly available through the CLARIN-LT Repository, Kaunas University of Technology. No new primary data were collected as part of this study.

Acknowledgments: The authors acknowledge the support of Vilnius Gediminas Technical University (VILNIUS TECH), Lithuania, for providing the enabling environment for this study to be conducted.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network

BiLSTM	Bidirectional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
XLM-RoBERTa	Cross-Lingual Language Model-RoBERTa
SVM	Support Vector Machine
LR	Logistic Regression
LLM	Large Language Model
XAI	Explainable Artificial Intelligence
LIME	Local Interpretable Model-Agnostic Explanations
ROC-AUC	Receiver Operating Characteristic–Area Under the Curve
Syn-Aug	Synonym Replacement Augmentation
XLM-Aug	XLM-RoBERTa-Based Augmentation
GPT-Aug	GPT-Based Augmentation

Appendix A. GPT-4o Prompt Template for Lithuanian Data Augmentation

System prompt:

You are assisting with academic data augmentation for a Lithuanian harmful-content classification study. You will receive one Lithuanian training example and its binary class label.

Produce exactly ONE new Lithuanian paraphrase of the supplied text.
Requirements:

1. Preserve the original semantic intent and the supplied class label.
2. Keep the text in Lithuanian.
3. Do not translate it into another language.
4. Do not add explanations, labels, quotation marks, numbering, or metadata.
5. Do not sanitize a harmful example into a harmless one and do not make a non-harmful example hateful.
6. Vary wording and phrasing while remaining close enough to the original meaning to function as a label-preserving augmentation.
7. Return only the augmented Lithuanian text.

User prompt template:

Class label: {label} ({label_name})
Original Lithuanian text:
{original_text}
Return one label-preserving Lithuanian augmented version.

References

1. Zhang, W.; Sun, Z.; Li, Y.; Dai, L.; Bilal, H.; Ayouni, S.; Gadekallu, T.R. Toward resilient Chinese cyberbullying detection: A pinyin and adversarial learning framework. *Appl. Soft Comput.* **2026**, *201*, 115633. [CrossRef]
2. Khot, P.S.; Gurav, U.; Pujari, J.; Naik, U. Natural language processing and cellular automata based fine-tuned deep learning model for cyberbullying detection. *Comput. Electr. Eng.* **2026**, *136*, 111208. [CrossRef]
3. Akar, F. Performance analysis of NLP-based machine learning algorithms in cyberbullying detection. *Erzincan Univ. J. Sci. Technol.* **2024**, *17*, 445–459. [CrossRef]
4. Mahmud, M.I.; Mamun, M.; Abdelgawad, A. A deep analysis of textual features based cyberbullying detection using machine learning. In Proceedings of the 2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT 2022), Alamein New City, Egypt, 18–21 December 2022; pp. 166–170. [CrossRef]
5. Gamal, D.; Alfonse, M.; Jiménez-Zafra, S.M.; Aref, M. Intelligent multi-lingual cyber-hate detection in online social networks: Taxonomy, approaches, datasets, and open challenges. *Big Data Cogn. Comput.* **2023**, *7*, 58. [CrossRef]
6. Kaur, S.; Singh, S.; Kaushal, S. Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data. *Int. J. Cogn. Comput. Eng.* **2024**, *5*, 104–122. [CrossRef]

7. Sathya, J.; Harin Fernandez, F.M. Effective automatic cyberbullying detection using a hybrid approach SVM and NLP. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6. [CrossRef]
8. Yoheswari, S. Optimized cyberbullying detection in social media using supervised machine learning and NLP techniques. *J. Sci. Technol. Res.* **2024**, *5*, 421–429.
9. Maity, K.; Jain, R.; Jha, P.; Saha, S. Explainable cyberbullying detection in Hinglish: A generative approach. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 3338–3347. [CrossRef]
10. Daraghmi, E.-Y.; Qadan, S.; Daraghmi, Y.-A.; Yousuf, R.; Cheikhrouhou, O.; Baz, M. From text to insight: An integrated CNN-BiLSTM-GRU model for Arabic cyberbullying detection. *IEEE Access* **2024**, *12*, 103504–103519. [CrossRef]
11. Balaji, P.G.; Katariya, P.P.; Sruthi, S.; Venugopalan, M. Cyberbullying detection on multiclass data using machine learning and a hybrid CNN-BiLSTM architecture. In Proceedings of the 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS 2024), Chikkaballapur, India, 18–19 April 2024; pp. 1–6. [CrossRef]
12. Daniel, R.; Murthy, T.S.; Kumari, C.D.V.P.; Lydia, E.L.; Ishak, M.K.; Hadjouni, M.; Mostafa, S.M. Ensemble learning with tournament selected glowworm swarm optimization algorithm for cyberbullying detection on social-media. *IEEE Access* **2023**, *11*, 123392–123400. [CrossRef]
13. Priyadharshini, M.; Hemagowri, J.; Banu, A.F.; Nithya, B.; K, S.; Murugesh, V. Advanced cyberbullying detection: A hybrid model integrated with Naïve Bayes. In Proceedings of the 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 24–26 July 2024; pp. 1–5. [CrossRef]
14. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. 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)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
15. Mahdi, M.A.; Fati, S.M.; Hazber, M.A.G.; Ahamad, S.; Saad, S.A. Enhancing Arabic cyberbullying detection with end-to-end transformer model. *Comput. Model. Eng. Sci.* **2024**, *141*, 1651–1671. [CrossRef]
16. León-Paredes, G.A.; Bravo-Quezada, O.G.; Bermeo-Aguaysa, P.P.; Peláez-Currillo, M.J.; Jiménez-González, L.L. Preventing cyberbullying on social networks with Spanish parental control NLP system. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 1380–1391. [CrossRef]
17. Mandravickaitė, J.; Rimkienė, E.; Petkevičius, M.; Songailaitė, M.; Zaranka, E.; Krilavičius, T. Exploring hate speech detection models for Lithuanian language. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 206–218.
18. Cao, J.; Xiong, Y.; Wang, W.; Chen, G. Cyberbullying detection based on hybrid neural networks and multi-feature fusion. *Information* **2026**, *17*, 205. [CrossRef]
19. Yerekesheva, M.; Akhmetova, O.; Kaziyeva, A.; Sultan, D.; Toktarova, A.; Abdrakhmanov, R.; Abdimukhan, T. A hybrid BERT-CNN with multihead self-attention for automated cyberbullying detection. *Eng. Technol. Appl. Sci. Res.* **2026**, *16*, 30716–30724. [CrossRef]
20. Uke, P.S. Bilingual cyber bullying detection system: Enhancing online safety. In Proceedings of the 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Kirtipur, Nepal, 3–5 October 2024; pp. 837–842. [CrossRef]
21. Kumar, D. A hybrid DeBERTa and gated broad learning system for cyberbullying detection in English text. *arXiv* **2025**, arXiv:2506.16052. [CrossRef]
22. Wen, R.; Crowe, S.; Gupta, K.; Li, X.; Billinghurst, M.; Hoermann, S.; Allan, D.D.; Nassani, A.; Piumsomboon, T. Large language models for automatic detection of sensitive topics. In *Proceedings of the 36th Australasian Conference on Human-Computer Interaction (OzCHI 2024)*; Association for Computing Machinery: New York, NY, USA, 2025. [CrossRef]
23. Cirillo, S.; Desiato, D.; Polese, G.; Solimando, G.; Sugumaran, V.; Sundaramurthy, S. Exploring the ability of emerging large language models to detect cyberbullying in social posts through new prompt-based classification approaches. *Inf. Process. Manag.* **2025**, *62*, 104043. [CrossRef]
24. Nirmal, A.; Bhattacharjee, A.; Sheth, P.; Liu, H. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 223–233. [CrossRef]
25. Calabrese, A.; Neves, L.; Shah, N.; Bos, M.; Ross, B.; Lapata, M.; Barbieri, F. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 398–408. [CrossRef]
26. Salles, I.; Vargas, F.; Benevenuto, F. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 6659–6669.

27. Hu, Y.; Lee, R.K.W. HateXScore: A Metric Suite for Evaluating Reasoning Quality in Hate Speech Explanations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; pp. 4221–4240. [CrossRef]
28. Wang, J.; Fu, K.; Lu, C.T. SosNet: A graph convolutional network approach to fine-grained cyberbullying detection. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data 2020), Atlanta, GA, USA, 10–13 December 2020; pp. 1699–1708. [CrossRef]
29. Butkienė, R.; Dambrauskas, E.; Šukys, A.; Žitkus, V. Lithuanian Hate Speech Corpus V.1; Kaunas University of Technology, CLARIN-LT Repository, 2025. Available online: https://hdl.handle.net/20.500.11821/69 (accessed on 15 March 2026).
30. Vanigotha, A.R.; Naveen Kumar, M.R.; Hiremath, S.; Adityan, S.S.; Basha, M.J. Effective cyberbullying detection with SparkNLP. *Int. J. Res. Appl. Sci. Eng. Technol.* **2023**, *11*, 101–106. [CrossRef]
31. Agrawal, S.; Awekar, A. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval: Proceedings of ECIR*; Springer: Cham, Switzerland, 2018; pp. 141–153. [CrossRef]
32. Bankar, S.; Janrao, S.; Gupta, P.; Patil, R.; Nandkar, M.; Jalote, A.; Khan, K.S.; Mathew, K.N. Cyberbullying detection on Twitter using sentiment analysis. *J. Electr. Syst.* **2024**, *20–23*, 3568–3576.
33. Guo, X.; Anjum, U.; Zhan, J. Cyberbully detection using BERT with augmented texts. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 1246–1253. [CrossRef]
34. Pendela, N.P.S.; Janet, K.A.; Yadav, A.M.R.; Subramanyam, C.B.; Hariharan, S.; Kekreja, V. Enhancing cyberbullying detection: A multi-algorithmic approach. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–5. [CrossRef]
35. DeYoung, J.; Jain, S.; Rajani, N.F.; Lehman, E.; Xiong, C.; Socher, R.; Wallace, B.C. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4443–4458.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.