Exploring Hate Speech Detection Models for Lithuanian Language

Justina Mandravickaitė and **Eglė Rimkienė** and **Mindaugas Petkevičius**
and **Milita Songailaitė** and **Eimantas Zaranka** and **Tomas Krilavičius**
Vytautas Magnus University, Kaunas, Lithuania
name.surname@vdu.lt

Abstract

Online hate speech poses a significant challenge, as it can incite violence and contribute to social polarisation. This study evaluates traditional machine learning, deep learning and large language models (LLMs) for Lithuanian hate speech detection, addressing class imbalance issue via data augmentation and resampling techniques. Our dataset included 27,358 user-generated comments, annotated into Neutral language (56%), Offensive language (29%) and Hate speech (15%). We trained BiLSTM, LSTM, CNN, SVM, and Random Forest models and fine-tuned Multilingual BERT, LitLat BERT, Electra, RWKV, ChatGPT, LT-Llama-2, and Gemma-2 models. Additionally, we pre-trained Electra for Lithuanian. Models were evaluated using accuracy and weighted F1-score. On the imbalanced dataset, LitLat BERT (0.76 weighted F1-score) and Multilingual BERT (0.73 weighted F1-score) performed best. Over-sampling further boosted weighted F1-scores, with Multilingual BERT (0.85) and LitLat BERT (0.84) outperforming other models. Over-sampling combined with augmentation provided the best overall results. Under-sampling led to performance declines and was less effective. Finally, fine-tuning LLMs improved their accuracy which highlighted the importance of fine-tuning for more specialized NLP tasks.

1 Introduction

Online hate speech poses significant challenges as social media platforms, forums other online spaces often contain hateful content that can incite violence (Garland et al., 2020; Schäfer et al., 2024), reinforce discrimination (Noorian et al., 2024; Ștefăniță and Buf, 2021) and contribute to social polarisation (Stukal et al., 2022). Manually identifying and removing such content is time-consuming, costly and often inconsistent due to the high volume of online interactions (Hansen et al., 2024). Also, context (Markov and Daelemans, 2022; Yu et al., 2022), intent, and linguistic nuances (Lu et al., 2023) further complicate the detection of hate speech, making it difficult to distinguish between harmful and non-harmful speech.

Machine learning and deep learning models can analyze large volumes of text, identify patterns associated with hate speech and improve via data-driven learning (Malik et al., 2024). This study evaluates traditional machine learning algorithms, deep learning architectures and large language models (LLMs) for the detection of hate speech in Lithuanian in a scenario, where data and computational resources were limited. We experimented with different dataset configurations to assess how these models handle class imbalance issue which often occurs in many NLP tasks, including hate speech detection (Casula and Tonelli, 2020; Reddy and Rajalakshmi, 2020). Therefore, we analyzed the impact of data augmentation and resampling techniques on model performance to get insights into how to improve the detection of hate speech in Lithuanian, despite the issue of data imbalance. Also, existing hate speech research and resources focus on English or similar large languages and represent other languages much less (Tonneau et al., 2024), which motivated our work as well.

Thus the rest of the paper is structured as follows: Section 2 in brief describes related work, Section 3 presents our data, Section 4 introduces models we used in our experiments, Section 3 specifies experimental setup, Section 6 reports results and Section 7 ends the paper with conclusions.

2 Related Work

Hate speech is characterized by language that targets individuals or groups based on attributes such as gender, race, ethnicity, and disability (Sachdeva et al., 2021). The spread of hate speech is often encouraged by societal biases and conflicts

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 206–218
August 1, 2025 ©2025 Association for Computational Linguistics

(Poletto et al., 2021). Despite extensive research, defining hate speech remains challenging as it is dynamic and includes a broad range of concepts, such as incitement, impoliteness, stigmatization, and cyberbullying (Tontodimamma et al., 2021; Culpeper, 2021; Kansok-Dusche et al., 2023). The interpretation of hate speech is further complicated by cultural, political, and legal differences across different regions, which makes it highly context-dependent (Paz et al., 2020; Alkomah and Ma, 2022; Lee et al., 2024).

Popular hate speech detection approaches rely on machine learning and NLP techniques (Mullah and Zainon, 2021; Jahan and Oussalah, 2021). While traditional methods like logistic regression are still used (e.g., as in Rohith and Amanullah (2024)), deep learning architectures, such as Long Short-Term Memory (LSTM), have demonstrated high accuracy (Mullah and Zainon, 2021; Fazil et al., 2023). Hybrid models that integrate machine learning and deep learning techniques have further improved detection in successfully capturing lexical and contextual features of hate speech (Mullah and Zainon, 2021; Rawat et al., 2024).

Graph Neural Networks (GNNs) have gained attention due to their improved contextual understanding of hate speech (Rawat et al., 2024; Wasi, 2024). Also, it has been shown that additional contextual information relevant to hate speech detection can be obtained from related NLP tasks, such as sentiment analysis and emotion classification (Awal et al., 2021; del Arco et al., 2021; Jafari et al., 2023). Moreover, optimizing self-supervised and supervised learning techniques have been used for improving model accuracy in dual contrastive learning frameworks (Lu et al., 2023).

Automated hate speech detection is challenging due to its dynamic nature (Halevy, 2023). Therefore, an increasing number of studies focus on multilingual hate speech detection, particularly in low-resource settings (Awal et al., 2023; Gutha et al., 2023; Saha et al., 2023). Cross-lingual transfer learning has been employed to leverage high-resource languages, such as English, for improving detection in languages with limited annotated data (Bigoulaeva et al., 2021; Zia et al., 2022; de Oliveira et al., 2023). Furthermore, zero-shot transfer learning has shown promise in handling linguistic variations without requiring labeled data (Pamungkas et al., 2021; Zia et al., 2022; Castillo-López et al., 2023). Additionally, meta-learning frameworks such as HateMAML have been proposed to enhance cross-lingual transfer performance in low-resource settings (Mozafari et al., 2022; Awal et al., 2023), to name just a few.

Beyond cross-lingual strategies, data-efficient learning techniques have been explored to improve detection, e.g., even minimal fine-tuning data in the target language can significantly improve classification accuracy (Röttger et al., 2022). Also, data augmentation (Venturott and Ciarelli, 2020; Casula and Tonelli, 2024), oversampling (Sanya and Suadaa, 2022; Mohamed et al., 2023) and resampling (Reddy et al., 2023), ensemble learning (Chen et al., 2021; Mohamed et al., 2023; Daouadi et al., 2024), cost-sensitive learning (Sreelakshmi et al., 2024), etc. strategies were applied for solving class-imbalance problem.

Besides, the use of additional datasets has improved model performance in bilingual hate speech detection (Shahi and Majchrzak, 2024). Also, privacy-preserving techniques, such as federated learning, have been included in hate speech detection models to protect user data (Gala et al., 2023). Given the prevalence of multimodal hate speech, integrating visual and textual features is also relevant for comprehensive moderation (Gandhi et al., 2024). For example, multimodal approaches that incorporated textual and visual elements have been introduced for analyzing hate speech in memes (Barceló et al., 2024).

As deep learning models often function as "black boxes", it raises concerns about their interpretability and decision-making process. Therefore, efforts to improve transparency and to increase trust in automated moderation systems have been made, as in MacAvaney et al. (2019) and Wasi (2024).

Research on Lithuanian hate speech detection is still developing. It has a strong focus on linguistic analysis, such as analyzing the features of abusive and hateful comments in Lithuanian news (Ruzaitė, 2018, 2021) or discussing the definition of *hate* based on its usage in texts (Župerka, 2021). Also, first attempts in developing a corpus for hate speech detection in Lithuanian has been reported in (Gvozdovaitė et al., 2020). In addition, there is some initial research on the application of deep learning models for automatic hate speech detection in Lithuanian, which is reported in (Kankevičiūtė, 2023; Kankevičiūtė et al., 2023a,b).

Despite all advancements, challenges persist in validation, such as bias in training datasets and model overfitting. Therefore, frameworks such as HateCheck have been introduced to improve evaluation across linguistic and contextual settings (Röttger et al., 2021). Considering the continuous changes in hate discourse, refining detection methodologies, including scenarios and languages with limited resources, retains its importance.

3 Data

To develop a hate speech detection solution for the Lithuanian language, an initial dataset of approximately 60,000 comments was collected from various Lithuanian news portals, including *15min.lt*, *alkas.lt*, and *delfi.lt*. Additionally, the dataset was supplemented with 226,776 comments from news portal *lrytas.lt*[1] and manually collected hate speech comments from various social media pages and news portals[2]. The sources were selected based on their popularity, accessibility of user-generated comments and differences in their audiences (e.g., *alkas.lt* tends to express nationalist perspectives.)

Recent comments were gathered according to the specific topics (LGBT+, gender issues, immigrants, etc.) that were the most prevalent in the hate speech data. These themes were extracted on the basis of an initial quantitative and qualitative analysis to ensure that contemporary patterns of hate speech, such as the latest vocabulary, were covered.

3.1 Annotation Process

A total of 27,358 comments were manually annotated by four annotators[3]. Every comments was annotated by at least two annotators. Personally identifiable information (PII), when it occurred in the user-generated comments, was anonymized. The annotation scheme consisted of three classes:

- **Neutral language** – general user-generated comments without offensive or hateful content (56% 15 317 of all annotated comments);
- **Offensive language** – comments containing strong or harmful expressions but lacking explicit hate speech (Chen et al., 2012) (28,6% or 7821 of all annotated comments);
- **Hate speech** – comments directed at individuals or groups based on protected characteristics such as gender, race, ethnicity, or religion (Sachdeva et al., 2021) (15,4% or 4220 of all annotated comments).

For annotation, guidelines were prepared with definitions of hate speech, offensive and neutral language as well as examples. Pre-annotation exercise was employed to guarantee that annotators correctly understood their task. The exercise consisted of small sample of user-generated comments which were annotated by all the annotators together with researchers, who prepared the guidelines, leading the activity.

At the first stage of annotation, every comment was annotated by two annotators. If they disagreed on the labels, the third annotator annotator was assigned to review the disagreement cases. The second stage included another review where disagreement cases were discussed by all the annotators until the agreement was reached. The user-generated comments for which the agreement was not reached even after second stage were not included in the final dataset.

Some comments contained racist or hateful content without explicit slurs, making the annotation process challenging as they needed additional discussions. Also, figurative and coded language required additional contextual knowledge for correct annotation. Finally, comments that contained only hyperlinks, names, symbols, or emojis were excluded from the final dataset.

Despite our efforts, our final dataset was imbalanced in terms of class distribution[4]. This imbalance presented challenges for model training, as the hate speech class had considerably fewer examples. To address this, multiple dataset balancing techniques were explored (see Section 5).

3.2 Data Augmentation

To address the effects of data imbalance and enhance model robustness, we applied **lexical-based data augmentation** (Jahan et al., 2024). This technique was applied to *hate speech* and *offensive speech* categories to artificially increase their representation:

- Keywords in hate speech and offensive comments were replaced with synonyms or stylistically similar terms while maintaining the original meaning.

[1] These comments were provided to us by *lrytas.lt* by agreement.

[2] In this stage, we applied targeted search and used the Google search engine to search for user-generated comments on Lithuanian news portals based on a set of keywords and key phrases. This set of keywords and key phrases were identified by exploring several random samples of our data quantitatively and qualitatively.

[3] The user-generated comments were sampled randomly and annotated in the period of two months.

[4] The dataset will be available upon request.

- Predefined lexicons[5] of interchangeable words were used to ensure context-aware modifications.
- Some offensive or hate-inducing words were substituted with alternative expressions that preserved negative connotations.

This approach helped to increase data diversity while ensuring that models were not overly sensitive to specific word choices. By introducing different variations of hate speech and offensive content, the models became more adaptable to subtle lexical changes in real-world data.

4 Models

For hate speech detection in Lithuanian, we employed well-known deep learning models, including Multilingual BERT, LitLat BERT, Electra, RWKV, ChatGPT, LT-Llama-2, Gemma-2, BiLSTM, LSTM, CNN, as well as traditional machine learning models such as SVM, and Random Forest. Lithuanian is considered a lesser-resourced language, therefore still not many pre-trained models can adequately process Lithuanian texts. Of these models, BiLSTM, LSTM, CNN, SVM, and Random Forest were trained for hate speech detection in Lithuanian, while Multilingual BERT, LitLat BERT, Electra, RWKV, ChatGPT, LT-Llama-2, Gemma-2 were further fine-tuned to classify Lithuanian user-generated comments, identifying those that may contain hate speech. Also, we pre-trained Electra model for Lithuanian from scratch.

Multilingual BERT. This model[6] uses the architecture of the BERT model and is trained on 104 languages, including Lithuanian (Pires et al., 2019). Wikipedia[7] texts were used to train this model.

LitLat BERT. It is a trilingual model[8] that was built using the *XLM-RoBERTa-base* (Zhao and Tao, 2021) and trained on Lithuanian, Latvian, and English language data.

Electra transformer. It is a transformer model[9] that uses a pre-training method which trains two neural network models: a generator and a discriminator. This proposed training method is significantly more efficient than the masked training method used in BERT models. This is why the *Electra* model requires fewer data and computer resources for training (Clark et al., 2020).

As there was no pre-trained *Electra* for Lithuanian, we pre-trained it ourselves.

RWKV. It is a language model that combines transformers and recurrent neural networks (RNNs) (Peng et al., 2023). RWKV can effectively use past context while avoiding some common challenges of traditional RNNs, such as struggling to handle long sequences. However, one of its main weaknesses is its sensitivity to how information is presented, i.e., reordering words in a prompt can significantly impact its performance.

ChatGPT. In our experiments, we used ChatGPT models via the OpenAI API. We applied a few-shot learning approach (Parnami and Lee, 2022) to expose the models to multiple labeled examples to improve classification consistency. We defined a structured system prompt for the hate speech detection task and provided representative samples to ensure that the models differentiate between examples of hate, offensive and neutral speech.

LT-Llama-2. We fine-tuned LT-Llama-2-7B-Instruct model (Nakvosas et al., 2024) using LoRA (Low-Rank Adaptation) (Hu et al., 2022). During the fine-tuning process, our dataset was pre-processed and formatted into instruction-based prompts. Fine-tuning used gradient accumulation, AdamW optimization, and early stopping.

Gemma-2. For hate speech detection, we fine-tuned the Gemma-2B model (Team et al., 2024) with LoRA and 4-bit quantization. The dataset was pre-processed and formatted into instruction-based prompts to ensure consistency during training. We employed gradient accumulation and the AdamW optimizer. Fine-tuning was monitored with early stopping to prevent overfitting.

BiLSTM and LSTM. A BiLSTM model (Cui et al., 2018) was trained using pre-trained FastText embeddings (Bojanowski et al., 2016). The text data underwent tokenization, padding, and encoding, with the model comprising an embedding layer, two BiLSTM layers (128 units each), and a softmax classifier. Sparse categorical cross-entropy

[5] These lexicons were developed based on the findings of initial quantitative and qualitative analysis of our corpus. They will be made public in the future.

[6] https://huggingface.co/google-bert/bert-base-multilingual-cased

[7] https://www.wikipedia.org/

[8] https://huggingface.co/EMBEDDIA/litlat-bert

[9] https://huggingface.co/docs/transformers/en/model_doc/electra

was used for training, with Adam optimizer and early stopping to prevent overfitting.

Similarly, an LSTM-based model was implemented with pre-trained FastText embeddings as well. The architecture consists of an embedding layer, two stacked LSTM layers (128 units each), and a softmax classifier. Sparse categorical cross-entropy and the Adam optimizer were applied, incorporating early stopping.

CNN. A model based on a convolutional neural network (CNN) (O'shea and Nash, 2015) was developed, again, with pre-trained FastText embeddings. The architecture includes an embedding layer, a 1D convolutional layer (128 filters, kernel size 5), a global max pooling layer, and fully connected layers with dropout for regularization. The training was, again, performed with sparse categorical cross-entropy and the Adam optimizer.

SVM. Using FastText sentence embeddings, we also trained a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) classifier. The dataset underwent preprocessing to handle missing values and filter valid labels. Each text sample was transformed into FastText embeddings before training an SVM model with a linear kernel.

Random Forest. The Random Forest (Ho, 1995) classifier was developed with, yet again, FastText sentence embeddings. Preprocessing included text transformation into FastText embeddings and label encoding for classification. The model was trained with 100 estimators.

4.1 Evaluation Methods

We evaluated our models using accuracy and weighted F1-score. Accuracy measures the proportion of correctly classified instances over the total number of instances. However, it can be misleading in cases of class imbalance, where the majority class dominates the predictions (Hort, 2023).

To provide a more balanced assessment, we used the weighted F1-score. This metric calculates the F1-score for each class independently and then averages them, weighting each class by its number of true instances (support). Unlike macro averaging, which treats all classes equally, the weighted F1-score ensures that classes with more samples contribute proportionally, making it particularly suitable for datasets with imbalanced class distributions (Vina, 2024).

5 Experimental Setup

Experimental setup includes dataset configurations model training and evaluation as well as a separate LLMs testing procedure. The code used for these experiments as well as model training parameters are available at `https://github.com/CARD-AI/LT-hatespeech-models`.

5.1 Dataset Configurations

Since our annotated corpus was not balanced across classes, we needed a way to get the best possible results with the available data. To address this issue, we explored different approaches to handling an imbalanced dataset for classifying task user-generated comments into neutral, offensive, or hate speech. We conducted three experiments using our data in different ways:

- **Dataset balanced with under-sampling** – the classes were balanced by reducing the number of offensive and neutral language comments to match the class with the fewest comments (hate speech)[10];
- **Dataset balanced with over-sampling** – the classes were balanced by increasing the number of hate speech and offensive language comments[11] to match the class with the highest number of comments (neutral language);
- **Original imbalanced dataset** – our original annotated dataset without any modification.

After applying lexical data augmentation (Section 3.2), we generated an additional augmented dataset. The same balancing techniques (under-sampling and over-sampling) were applied to the augmented data, allowing us to analyze whether augmentation improves classification performance.

5.2 Model Training and Evaluation

For each dataset configuration, we trained and evaluated the following models:

- **Traditional deep learning models:** BiLSTM, LSTM, CNN.

[10] The examples to remove were selected randomly. After the under-sampling was applied, the number of offensive and neutral language comments each was equal to the number of hate speech comments.

[11] For balancing dataset with oversampling, the number of hate speech and offensive language comments was increased by making copies of them. The comments were selected form these classes randomly and the copies were made until the number of comments for each of these two classes was equal to the number of comments of the neutral language.

- **Machine learning classifiers:** SVM, Random Forest.
- **LLMs:** LT-Llama-2 (LT-Llama-2-7B & LT-Llama-13b), Gemma-2 (Gemma-2-2b & Gemma-2-9b), and OpenAI's ChatGPT models (GPT-4o, GPT-4o-mini, GPT-4, GPT-3.5-Turbo).
- **Other transformer-based models:** Multilingual BERT, LitLat BERT, Electra, RWKV.

For experiments, 80 % of the dataset was used for training, 10 % – for validation and 10 % – for testing (the exceptions are LLMs; their testing procedure is described in 5.2.1.). Each model was evaluated using **accuracy** and **weighted F1-score**.

5.2.1 LLMs Testing Procedure

To evaluate the impact of fine-tuning, we tested both pre-trained and fine-tuned versions of LLMs that we used in our experiments. The pre-trained models included two versions of Gemma-2 (Gemma-2-2b & Gemma-2-9b), LT-Llama-2-13B, and OpenAI's ChatGPT models. Additionally, we fine-tuned a smaller Gemma-2 model and an LT-Llama-2-7B model on our dataset.

Pre-trained models were evaluated on the mini testing dataset, which contains 50 comments per class. We made this decision in order to mitigate computational cost, which was especially relevant in terms of using OpenAI's ChatGPT models. Therefore, as the results for pre-trained and fine-tuned LLMs were presented in separate subsection as they were not directly comparable with the results of our other experiments. Fine-tuned models were trained on the original dataset before evaluation. To balance performance and computational efficiency, we also applied appropriate model quantization settings.

6 Results

This section presents the evaluation results of different hate speech detection models trained and tested on Lithuanian data. The models were assessed across 3 dataset variants: the original imbalanced dataset, a dataset balanced with over-sampling and a dataset balanced with under-sampling. Additionally, an augmented dataset was evaluated using the same data balancing strategies to determine whether it could improve model performance. We used accuracy and weighted F1-score as the primary evaluation metrics.

6.1 Model Performance with the Imbalanced Original Dataset

Table 1 presents model performance on the original imbalanced dataset. The best-performing models were LitLat BERT (0.76 accuracy, 0.76 weighted F1-score) and Multilingual BERT (0.73 accuracy, 0.73 weighted F1-score), demonstrating that transformer-based architectures fine-tuned for multilingual or Lithuanian texts excel in this task.

Traditional machine learning models like SVM and Random Forest performed moderately (0.66 and 0.63 accuracy, respectively), showing that while they capture useful text representations, they struggle with nuanced language complexities. Deep learning architectures such as CNN, LSTM, and BiLSTM exhibited slightly lower performance, with BiLSTM achieving the best result (0.68 accuracy, 0.65 weighted F1-score). RWKV performed the worst (0.57 accuracy, 0.49 weighted F1-score), suggesting its recurrent-based approach is less suited for this task.

Model	Accuracy	Weighted F1-score
CNN	0.67	0.62
LSTM	0.66	0.61
BiLSTM	0.68	0.65
RWKV	0.57	0.49
SVM	0.66	0.56
Random Forest	0.63	0.51
LitLat BERT	**0.76**	**0.76**
Electra	0.73	0.72
Multilingual BERT	0.73	0.73

Table 1: Model Performance on Imbalanced Dataset (Baseline Results)

To streamline the analysis and highlight the most relevant findings, subsequent tables (Tables 2, 4, and 6) include only a subset of models. Specifically, we focus on transformer-based architectures (LitLat BERT, Electra, and Multilingual BERT) which consistently demonstrated the highest performance in the initial evaluation (Table 1). These models are the most promising for practical applications due to their robustness and multilingual capabilities. RWKV is included as an experimental baseline due to its novel recurrent-based architecture, which we were interested in benchmarking against transformers, despite its overall lower performance.

After applying augmentation techniques, the accuracy of all models improved (Table 2). Notably, LitLat BERT's accuracy increased to 0.83, and Multilingual BERT rose to 0.76 accuracy, indicating that augmented datasets enhance feature representation. Electra and RWKV also improved (Electra: 0.72→0.73 accuracy; RWKV: 0.57→0.58), though RWKV remained weaker than transformer-based models.

Model	**Accuracy**	**Weighted F1-score**
RWKV	0.58	0.54
LitLat BERT	**0.83**	**0.83**
Electra	0.72	0.69
Multilingual BERT	0.76	0.76

Table 2: Impact of Data Augmentation on Model Performance

6.2 Model Performance with the Dataset Balanced via Over-Sampling

Balancing the dataset through over-sampling led to minor improvements in accuracy for some models, while others showed little change or slightly decreased performance (Table 3). Multilingual BERT (0.85 accuracy, 0.85 weighted F1-score) and LitLat BERT (0.84 accuracy, 0.84 weighted F1-score) outperformed all other models. CNN, LSTM, and BiLSTM had modest improvements, while RWKV improved slightly but remained behind transformers.

Over-sampling mitigates class imbalance but may introduce redundancy, leading to overfitting, particularly for traditional models like Random Forest (0.65 accuracy, 0.54 weighted F1-score) and SVM (0.65 accuracy, 0.62 weighted F1-score).

Table 4 presents results for models trained on the dataset augmented dataset, balanced via over-sampling. Multilingual BERT emerged as the top performer in this setup, reaching 0.85 in both accuracy and weighted F1-score. LitLat BERT followed closely with 0.84 accuracy, demonstrating that it also benefited from the balancing techniques. Electra saw moderate improvements, achieving 0.75 accuracy. Meanwhile, RWKV continued to lag behind, indicating that its recurrent-based approach struggled with this task. The findings suggest that over-sampling effectively mitigates class imbalance and, when combined with augmentation, leads to significant performance gains, particularly for transformer-based models.

Model	**Accuracy**	**Weighted F1-score**
CNN	0.66	0.61
LSTM	0.61	0.55
BiLSTM	0.62	0.61
RWKV	0.56	0.56
SVM	0.65	0.62
Random Forest	0.65	0.54
LitLat BERT	**0.76**	**0.77**
Electra	0.73	0.72
Multilingual BERT	0.72	0.72

Table 3: Model Performance with Dataset Balanced via Over-Sampling

Model	**Accuracy**	**Weighted F1-score**
RWKV	0.56	0.56
LitLat BERT	0.84	0.84
Electra	0.75	0.76
Multilingual BERT	**0.85**	**0.85**

Table 4: Effect of Over-Sampling and Data Augmentation on Model Performance

6.3 Model Performance with the Dataset Balanced via Under-Sampling

Reducing neutral and offensive language classes to match the number of hate speech instances resulted in decreased performance across models (Table 5). LitLat BERT (0.74 accuracy, 0.74 weighted F1-score) and Electra (0.69 accuracy, 0.69 weighted F1-score) were the top performers, but overall accuracy dropped compared to over-sampling.

Traditional classifiers, SVM and Random Forest, dropped significantly (SVM: 0.64 accuracy, Random Forest: 0.56 accuracy), highlighting their dependency on larger training datasets. RWKV struggled (0.40 accuracy, 0.41 weighted F1-score), reaffirming its weaker performance in hate speech detection.

Table 6 demonstrates the effect of under-sampling alongside augmentation. LitLat BERT and Multilingual BERT achieved identical accuracy scores of 0.76, showing that even with fewer training examples, augmentation provided a performance boost. Electra, however, saw a slight drop in accuracy to 0.68, suggesting that it relies more on a larger dataset to generalize well. RWKV improved from 0.40 to 0.53 accuracy, though it still under-

Model	Accuracy	Weighted F1-score
CNN	0.60	0.58
LSTM	0.62	0.59
BiLSTM	0.62	0.61
RWKV	0.40	0.41
SVM	0.64	0.62
Random Forest	0.56	0.52
LitLat BERT	**0.74**	**0.74**
Electra	0.69	0.69
Multilingual BERT	0.68	0.68

Table 5: Model Performance with Dataset Balanced via Under-Sampling

performed compared to transformer-based models. While under-sampling with augmentation was beneficial, it proved less effective than over-sampling approaches.

Model	Accuracy	Weighted F1-score
RWKV	0.53	0.52
LitLat BERT	**0.76**	**0.77**
Electra	0.68	0.67
Multilingual BERT	0.76	0.75

Table 6: Effect of Under-Sampling and Data Augmentation on Model Performance

6.4 Model Performance Comparison of Pre-Trained and Fine-Tuned LLMs

This section presents a comparison between pre-trained and fine-tuned LLMs using a common test dataset in order to explore the effect of fine-tuning for model performance. Table 7 summarizes the accuracy and weighted F1-scores for each model.

The results indicate that fine-tuning improves classification accuracy for LLMs in hate speech detection (see Table 7). The *LT-Llama-2-7B fine-tuned model* achieved the highest accuracy (0.74) and weighted F1-score (0.75), demonstrating the benefits of training on domain-specific data. Meanwhile, *Chatgpt-gpt4o*, second best result, reached 0.7 accuracy and weighted F1-score without fine-tuning. However, it is significantly larger than *LT-Llama-2-7B* (although OpenAI has not officially disclosed the exact number of parameters, it has been reported that Chatgpt-gpt4o may be over one trillion parameters (Shahriar et al., 2024).).

Model	Accuracy	Weighted F1-score
Gemma-2-2b-it-Q8-0.gguf	0.47	0.46
Gemma-2-9b-it-Q6-K-L.gguf	0.59	0.58
Chatgpt-gpt4o	0.70	0.70
Chatgpt-gpt4o-mini	0.69	0.69
Chatgpt-gpt4	0.69	0.69
Chatgpt-gpt3.5-turbo	0.55	0.51
LT-Llama-2-13b-q8	0.16	0.13
Gemma-2-2b-it-finetuned	0.56	0.55
LT-Llama-2-7B-finetuned	**0.74**	**0.75**

Table 7: Performance Comparison of Pre-Trained and Fine-Tuned Models)

A direct comparison of pre-trained and fine-tuned versions of the *Gemma-2-2B* model shows an increase in accuracy from 0.47 to 0.56, suggesting that even relatively small-scale fine-tuning can lead to performance gains. Similarly, *LT-Llama-2-13B-q8*, which performed poorly (0.16 accuracy), highlights the limitations of using a model without fine-tuning for specialized classification tasks.

Meanwhile, *GPT-3.5-turbo* lagged behind other tested OpenAI models with 0.55 accuracy and 0.51 weighted F-1 score, which suggests that GPT-4's improvements matter for classification. Also, *Gemma-2-9B-it-Q6-K-L.gguf* (pre-trained) outperformed *GPT-3.5-turbo*, indicating that certain smaller architectures can outperform larger LLMs, while fine-tuning can boost performance further.

These findings align with existing research, such as in Sen et al. (2024) and Wullach et al. (2021), which suggests that fine-tuned models outperform generic pre-trained ones in domain-specific tasks like hate speech detection.

7 Conclusions

This study evaluates traditional machine learning algorithms, deep learning architectures and LLMs in detecting hate speech in Lithuanian in a scenario where data and computational resources were limited. We experimented with different dataset configurations to assess how these models handle class imbalance issue. Therefore, we analyzed the impact of data augmentation via lexical substitution and resampling techniques on model performance.

Transformer-based models, particularly LitLat BERT and Electra, demonstrated the best performance on the imbalanced original dataset, achieving 0.76 and 0.73 weighted F-1 scores, respectively. Traditional machine learning models like SVM and Random Forest performed moderately, while RWKV struggled (0.49 weighted F-1 score). Data augmentation improved all models, with LitLat BERT reaching 0.83 weighted F-1 score.

With an over-sampling, accuracy improved further. Multilingual BERT (0.85 weighted F-1 score) and LitLat BERT (0.84 weighted F-1 score) outperformed all other models. Electra improved to 0.76 weighted F-1 score, while RWKV remained the weakest. Traditional models like SVM (0.62 weighted F-1 score) and Random Forest (0.54 weighted F-1 score) suffered from potential overfitting. Over-sampling combined with augmentation provided the best results, particularly for transformer-based models.

With an under-sampling, performance declined. LitLat BERT (0.74 weighted F-1 score) and Electra (0.69 weighted F-1 score) still performed well, but SVM (0.62 weighted F-1 score) and Random Forest (0.52 weighted F-1 score) dropped significantly. RWKV struggled with 0.41 weighted F-1 score, though augmentation slightly improved it. Undersampling was less effective than over-sampling for performance gains.

GPT-4 models performed well enough without fine-tuning but at higher computational and financial cost. Fine-tuning smaller LLMs significantly boosted their accuracy. Fine-tuned *LT-Llama-2-7B* achieved 0.75 weighted F-1 score, while *Gemma-2-2B-it* improved from 0.46 to 0.55 weighted F-1 score. However, LT-Llama-2-13B-q8 (0.16 accuracy) showed that pre-trained models without fine-tuning perform poorly on specialized tasks.

Our future plans include increasing our dataset as well as exploring larger variety of models and fine-tuning techniques. We also plan to examine model biases and decision-making processes.

Limitations

The effectiveness of hate speech detection in our experiments has been constrained by our dataset limitations, as its scope may not adequately reflect the diversity of online discourse. Also, sampling biases and performed anonymization might have obscured contextual cues, which could have affected model accuracy. Additionally, context-dependent hate speech variations may be insufficiently represented, which reduces applicability of our results to real-world scenarios.

Furthermore, annotation bias may have further complicated hate speech detection, as subjective interpretations by human annotators influence dataset labels. Also, the dynamic nature of hate speech introduces new expressions that static datasets do not capture, thus needing constant updates.

Challenges arise from data balancing techniques as well, which may introduce biases that impact model robustness. In addition, deep learning models, LLMs included, lack proper explainability, which limits their transparency in decision-making processes.

Finally, automatic evaluation metrics are insufficient for capturing the nuanced and context-dependent nature of hate speech, therefore our experiments would benefit from a more comprehensive evaluation procedure, including comprehensive error analysis which we plan to include in our future research.

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 701–713. Springer.

Rabiul Awal, R. Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. 2023. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions on Computational Social Systems*, 11:1086–1095.

Sofía Barceló, Magalí Boulanger, Antonela Tommasel, and Juan Manuel Rodriguez. 2024. Beyond words: A preliminary study for multimodal hate speech detection. In *2024 L Latin American Computer Conference (CLEI)*, pages 1–4. IEEE.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the first workshop on language technology for equality, diversity and inclusion*, pages 15–25.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Galo Castillo-López, Arij Riabi, and Djamé Seddah. 2023. Analyzing zero-shot transfer scenarios across spanish variants for hate speech detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13.

Camilla Casula and Sara Tonelli. 2020. Hate speech detection with machine-translated data: The role of annotation scheme, class imbalance and undersampling. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*.

Camilla Casula and Sara Tonelli. 2024. A target-aware analysis of data augmentation for hate speech detection. *arXiv preprint arXiv:2410.08053*.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 international conference on privacy, security, risk and trust and 2012 international confernece on social computing*, pages 71–80. Ieee.

Zhi Chen, Jiang Duan, Li Kang, and G. Qiu. 2021. Class-imbalanced deep learning via a class-balanced ensemble. *IEEE Transactions on Neural Networks and Learning Systems*, 33:5626–5640.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Zhiyong Cui, Ruimin Ke, Ziyuan Pu, and Yinhai Wang. 2018. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*.

Jonathan Culpeper. 2021. Impoliteness and hate speech: Compare and contrast. *Journal of pragmatics*, 179:4–11.

Kheir Eddine Daouadi, Yaakoub Boualleg, and Kheir Eddine Haouaouchi. 2024. Ensemble of pretrained language models and data augmentation for hate speech detection from arabic tweets. *arXiv preprint arXiv:2407.02448*.

Aillkeen Bezerra de Oliveira, Cláudio de Souza Baptista, Anderson Almeida Firmino, and Anselmo Cardoso de Paiva. 2023. Using multilingual approach in cross-lingual transfer learning to improve hate speech detection. In *ICEIS (1)*, pages 374–384.

Flor Miriam Plaza del Arco, M. Molina-González, L. A. U. López, and M. Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Mohd Fazil, Shakir Khan, Bader M Albahlal, Reemiah Muneer Alotaibi, Tamanna Siddiqui, and Mohd Asif Shah. 2023. Attentional multi-channel convolution with bidirectional lstm cell toward hate speech prediction. *IEEE Access*, 11:16801–16811.

Jay Gala, Deep Gandhi, Jash Mehta, and Zeerak Talat. 2023. A federated approach for hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3248–3259.

Ankita Gandhi, Param Ahir, K. Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and A. Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112.

Abhinav Reddy Gutha, Nidamanuri Sai Adarsh, Ananya Alekar, and Dinesh Reddy. 2023. Multilingual hate speech and offensive language detection of low resource languages. In *FIRE (Working Notes)*, pages 445–458.

Veronika Gvozdovaitė, Aušrinė Naujalytė, Justina Mandravickaitė, and Tomas Krilavičius. 2020. An overview of the lithuanian hate speech corpus. *Int. J. Des. Anal. Tools Integr. Circuits Syst*, pages 54–57.

Karina Halevy. 2023. A group-specific approach to nlp for hate speech detection. *arXiv preprint arXiv:2304.11223*.

T. M. Hansen, Lasse Lindekilde, S. Karg, Michael Bang Petersen, and S. Rasmussen. 2024. Combatting online hate: Crowd moderation and the public goods problem. *Communications*.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Max Hort. 2023. *Investigating trade-offs for fair machine learning systems*. Ph.D. thesis, UCL (University College London).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

A. Jafari, Guanlin Li, P. Rajapaksha, R. Farahbakhsh, and Noel Crespi. 2023. Fine-grained emotions influence on implicit hate speech detection. *IEEE Access*, 11:105330–105343.

Md Saroar Jahan and M. Oussalah. 2021. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Md Saroar Jahan, Mourad Oussalah, Djamila Romaissa Beddia, Nabil Arhab, et al. 2024. A comprehensive study on nlp data augmentation for hate speech detection: Legacy methods, bert, and llms. *arXiv preprint arXiv:2404.00303*.

Eglė Kankevičiūtė. 2023. Neapykantos kalbos atpažinimas panaudojant dirbtinį intelektą. Master's thesis, Vytautas Magnus University, Kaunas, Lithuania.

Eglė Kankevičiūtė, Milita Songailaitė, and Justina Mandravickaitė. 2023a. Neapykantos kalbos atpainimas lietuvikuose komentaruose panaudojant dirbtinį intelektą. *Vilnius University Open Series*.

Eglė Kankevičiūtė, Milita Songailaitė, Bohdan Zhyhun, and Justina Mandravickaitė. 2023b. Lithuanian hate speech classification using deep learning methods. *Automation of technological and business processes*.

Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, Anke Zeißig, Lisanne Seemann-Herz, Sebastian Wachs, and Ludwig Bilz. 2023. A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena. *Trauma, violence, & abuse*, 24(4):2598–2615.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224.

Junyu Lu, Ho-Yi Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and Bo Xu. 2023. Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2787–2795.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and O. Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS ONE*, 14.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2024. Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, pages 1–16.

Ilia Markov and Walter Daelemans. 2022. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42.

Mohamed S. Mohamed, Hossam Elzayady, K. Badran, and G. Salama. 2023. An efficient approach for data-imbalanced hate speech detection in arabic social media. *J. Intell. Fuzzy Syst.*, 45:6381–6390.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Nanlir Sallau Mullah and W. Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: A review. *IEEE Access*, 9:88364–88376.

Artūras Nakvosas, Povilas Daniušis, and Vytas Mulevičius. 2024. Open llama2 model for the lithuanian language. *arXiv preprint arXiv:2408.12963*.

Zeinab Noorian, Amira Ghenai, Hadiseh Moradisani, Fattane Zarrinkalam, and Soroush Zamani Alavijeh. 2024. User-centric modeling of online hate through the lens of psycholinguistic patterns and behaviors in social media. *IEEE Transactions on Computational Social Systems*, 11:4354–4366.

Keiron O'shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Endang Wahyu Pamungkas, Valerio Basile, and V. Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Inf. Process. Manag.*, 58:102544.

Archit Parnami and Minwoo Lee. 2022. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*.

María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. Rwkv: Reinventing rnns for the transformer era. *Preprint*, arXiv:2305.13048.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16.

B. A. C. Reddy, Girish Kumar Chandra, Dilip Singh Sisodia, and Arti Anuragi. 2023. Balancing techniques for improving automated detection of hate speech and offensive language on social media. *2023 2nd International Conference for Innovation in Technology (INOCON)*, pages 1–8.

Yashwanth Reddy and Ratnavel Rajalakshmi. 2020. Dlrg@ hasoc 2020: A hybrid approach for hate and offensive content identification in multilingual tweets. In *FIRE (working notes)*, pages 304–310.

Yalam Venkata Sai Rohith and M. Amanullah. 2024. Improving the accuracy by comparing the gaussian naive bayes algorithm and logistic regression for predicting hate speech recognition. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.

Jūratė Ruzaitė. 2018. In search of hate speech in lithuanian public discourse: A corpus-assisted analysis of online comments. *Lodz Papers in Pragmatics*, 14:116 – 93.

Jūratė Ruzaitė. 2021. How do haters hate? verbal aggression in lithuanian online comments. *Discourse and Conflict*.

J. Sachdeva, Kushank Kumar Chaudhary, Harshit Madaan, and P. Meel. 2021. Text based hate-speech analysis. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 661–668.

Sougata Saha, Michael Sullivan, and Rohini K Srihari. 2023. Hate speech detection in low resource indo-aryan languages. In *FIRE (Working Notes)*.

Almira Diva Sanya and L. H. Suadaa. 2022. Handling imbalanced dataset on hate speech detection in indonesian online news comments. *2022 10th International Conference on Information and Communication Technology (ICoICT)*, pages 380–385.

Svenja Schäfer, Isabella Rebasso, Ming Manuel Boyer, and Anna Maria Planitzer. 2024. Can we counteract hate? effects of online hate speech and counter speech on the perception of social groups. *Communication Research*, 51(5):553–579.

Tanmay Sen, Ansuman Das, and Mrinmay Sen. 2024. Hatetinyllm: hate speech detection using tiny large language models. *arXiv preprint arXiv:2405.01577*.

Gautam Kishore Shahi and Tim A. Majchrzak. 2024. Hate speech detection using cross-platform social media data in english and german language. In *International Conference on Web Information Systems and Technologies*.

Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Denis K Stukal, Andrei S Akhremenko, and Alexander PC Petrov. 2022. Affective political polarization and hate speech: Made for each other? *RUDN Journal of Political Science*, 24(3):480–498.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126:157–179.

L. Venturott and P. Ciarelli. 2020. Data augmentation for improving hate speech detection on social networks. *Proceedings of the Brazilian Symposium on Multimedia and the Web*.

Abirami Vina. 2024. What is f1 score? a computer vision guide. *Roboflow Blog*.

Azmine Toushik Wasi. 2024. Explainable identification of hate speech towards islam using graph neural networks. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 250–257.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. *arXiv preprint arXiv:2109.00591*.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930.

Yingjia Zhao and Xin Tao. 2021. Zyj123@ dravidianlangtech-eacl2021: Offensive language identification based on xlm-roberta with dpcnn. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 216–221.

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. In *Proceedings of the International AAAI conference on web and social media*, volume 16, pages 1435–1439.

Kazimieras Romualdas Župerka. 2021. Ar įmanoma apibrėti vadinamosios neapykantos kalbos ribas? *Acta humanitarica academiae Saulensis*.

Oana Ștefăniță and Diana-Maria Buf. 2021. Hate speech in social media and its effects on the lgbt community: A review of the current research. *Romanian Journal of Communication and Public Relations*, 23(1):47–55.