Improving Hate Speech Classification with Cross-Taxonomy Dataset Integration

Jan Fillies[1,2] **and Adrian Paschke**[1,2,3]
[1]Institut für Angewandte Informatik, Leipzig, Germany
[2]Freie Universität Berlin, Berlin, Germany
[3]Fraunhofer-Institut für Offene Kommunikationssysteme, Berlin, Germany
jan.fillies@fu-berlin.de

Abstract

Algorithmic hate speech detection faces significant challenges due to the diverse definitions and datasets used in research and practice. Social media platforms, legal frameworks, and institutions each apply distinct yet overlapping definitions, complicating classification efforts. This study addresses these challenges by demonstrating that existing datasets and taxonomies can be integrated into a unified model, enhancing prediction performance and reducing reliance on multiple specialized classifiers. The work introduces a universal taxonomy and a hate speech classifier capable of detecting a wide range of definitions within a single framework. Our approach is validated by combining two widely used but differently annotated datasets, showing improved classification performance on an independent test set. This work highlights the potential of dataset and taxonomy integration in advancing hate speech detection, increasing efficiency, and ensuring broader applicability across contexts.

1 Introduction

Research has shown a direct link between the rise of online hate speech and offline events (Lupu et al., 2023), highlighting the growing impact of digital platforms on real-world occurrences. As of April 2023, there are an estimated 4.8 billion global social media users, making up about 59.9% of the world's population (Kemp, 2023). This massive reach underscores the scale of the problem, with Facebook alone removing 38.3 million instances of hate speech in the first three quarters of 2023 (Dixon, 2023). These numbers emphasize both the urgency and magnitude of the issue, making it a top priority for the research community. The challenge lies in balancing the preservation of free speech with the need to protect individuals from harm. While algorithms play a key role in addressing this issue, they are just one part of a broader, multifaceted approach. In this context, this research aims to develop efficient and effective algorithmic solutions for hate speech detection.

One main challenge in the field is that the understanding of hate speech varies and is influenced by factors such as topic (Wiegand et al., 2019), author (Nejadgholi and Kiritchenko, 2020), and time (Justen et al., 2022), among others. Even within the legal context, it is a complex process deciding whether a statement should be classified as hateful or not. In response, research, private, and public entities have developed their own definitions and community standards, legal frameworks, or annotation guidelines (MacAvaney et al., 2019).

Especially in the research field, the available datasets heavily depend on the annotation procedure and the definitions of hate speech provided to the annotators (Vidgen and Derczynski, 2020). This dependence and wide variety of definitions makes it challenging to compare (Fortuna and Nunes, 2018) or merge datasets annotated within different annotation schemas. While the field of available annotated hate speech corpora is limited to begin with, this additional limitation of incompatibility further complicates efforts to provide general and reliable hate speech detection.

This research addresses this gap by providing a machine learning structure that combines existing definitions and datasets. It identifies mismatches in definitions, faults during the annotation combining process, and missing labels in datasets. The study demonstrates the feasibility of merging annotation schemas and datasets to detect a wider variety of hate speech definitions using just one trained classifier. It establishes that a single general taxonomy can be created and employed for multi-label federated training of a classifier, thereby improving prediction quality.

The approach is evaluated using two standard research datasets and their respective definitions. The outcome involves the creation of a comprehensive hate speech taxonomy and the training of a

Proceedings of LaTeCH-CLfL 2025, pages 148–159

general hate speech classifier.

The scripts used for preprocessing, dataset construction, training, and evaluation are available as part of the paper.[1] This offers a deeper insight and facilitates the reproducibility of our work. Please note that the used datasets have to be obtained from the cited sources.

2 Related Work

Datasets - The field of hate speech datasets is rapidly growing. Established datasets include (Hosseinmardi et al., 2015; de Gibert et al., 2018; ElSherief et al., 2018), while newer, smaller datasets (Fillies et al., 2023b, 2025) continue to emerge. A comprehensive overview is provided by Vidgen and Derczynski (2020). Analysis of these datasets highlights diverse annotation schemes (Chung et al., 2019), from binary labels to multi-class hierarchies (Ranasinghe and Zampieri, 2020). Universal annotation frameworks are also recognized (Bartalesi et al., 2006). However, no single benchmark dataset or universally accepted definition of hate speech exists (MacAvaney et al., 2019). The wide range of definitions has been extensively studied by Stephan (2020).

Algorithmic Detection - For detecting hate speech, toxic speech, abusive language, and related areas, the predominant algorithmic approach has utilized supervised transformer-based architectures (Mozafari et al., 2020; Poletto et al., 2021; Plaza-del Arco et al., 2023). Fine-tuning transformer models, particularly BERT (Devlin et al., 2019), has demonstrated significant performance enhancements compared to other methods (Liu et al., 2019a; Kirk et al., 2022; Fillies et al., 2023a). Recently, the focus has shifted towards using pre-trained large language models combined with prompting techniques for hate speech detection (Kim et al., 2023; Plaza-del Arco et al., 2023; Fillies and Paschke, 2024).

Taxonomy and Ontology Matching - Several researchers have aimed to create general hate speech ontologies (Stranisci et al., 2022; Sharma et al., 2018) and taxonomies (Salminen et al., 2018; Zufall et al., 2022; Lewandowska-Tomaszczyk et al., 2023). Salminen et al. (2018) integrated their taxonomy into a transformer-based hate speech detection model, partially building on existing taxonomies and combining them to annotate a new dataset. The practice of merging ontologies is well established (Shvaiko and Euzenat, 2013). However, no research has yet combined hate speech taxonomies to make existing datasets suitable for iterative federated learning.

Federated and Continuous Learning - Federated learning for hate speech detection is crucial as it mitigates privacy concerns related to data sharing. A key development is Zampieri et al. (2024), which introduces a binary hate speech classifier using a decentralized architecture, demonstrating superior performance across datasets while preserving privacy. Another significant study, Gala et al. (2023), explores multi-class federated learning on a static dataset with uniform annotations, disregarding annotation mismatches and emphasizing distributed training benefits. In continuous learning, Omrani et al. (2023) propose a novel framework for detecting problematic content by integrating various datasets and treating each label as an independent classification task.

This research directly builds upon the work of Zampieri et al. (2024), Gala et al. (2023), and Omrani et al. (2023). It extends the findings of Zampieri et al. (2024) and Gala et al. (2023) by demonstrating that federated training for hate speech detection is feasible not only for binary classification but also for multi-label hate speech datasets with varying definitions of hate speech. In relation to Omrani et al. (2023), it advances the research by integrating labels into a unified taxonomy with hierarchical aspects, introducing a deeper semantic relationship model, and showing that this model can be continuously adapted.

3 Methodology

The research is divided into three main parts. First, a general hate speech taxonomy is created. Second, this taxonomy is used to fine-tune a pre-trained multi-label hate speech detection model multiple times on two different datasets (see Sections 4 and 6). Lastly, continuous evaluation is conducted after each training cycle. Each step is detailed in this section (see Figure 1), with selected datasets, taxonomies, and the models serving as examples to demonstrate the approach's functionality.

1. In the first step, the taxonomies are combined into one general taxonomy. Here, the general taxonomy should include all the classes proposed by the underlying concepts. A class hierarchy is introduced to represent and adjust to

[1] https://github.com/fillies/HateSpeechCrossTaxonomyDatasetIntegration

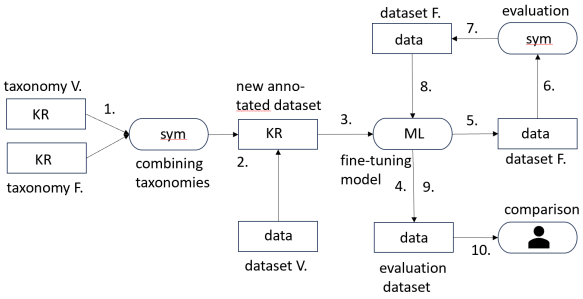

Figure 1: Boxology-Model of the Process. F. = Fanton et al. 2021, V. = Vidgen et al. 2021, sym = Symbolic Processing KR. = Knowledge Representation, ML = Machine Learning

different levels of abstraction (see section 5). In this step, classes that cannot be merged are identified and removed. A word-level matching of annotations between the original and the new general taxonomy is introduced. The class hierarchy of the general taxonomy is represented through a one-hot encoded vector; when a subclass is flagged as identified, the parent classes must be present too.

2. In the following step, one dataset is selected to have its annotations transferred into the new annotation format based on the general taxonomy. Here, it is expected that certain flags within the annotations are missing or, more precisely, incorrectly annotated.

3. Based on this newly annotated dataset, a multi-label classifier is trained (see section 6).

4. To validate the performance of the trained model and provide insight into the generalizability of the model, an external binary hate speech dataset is provided as an evaluation dataset, and the performance is measured (see section 6.6).

5. The trained classifier is now used to predict all known labels of the second dataset.

6. The True Positive, False Negative, False Positive, and True Negative distributions of the predictions generate insights into three main aspects regarding the annotations. Firstly, it can be observed where the definitions of concepts are not aligned. Secondly, it can be determined if the general taxonomy made a mistake in its hierarchical structure. Lastly, it can be identified which flags are not represented in the old annotation of the new dataset (see section 6.7).

7. After evaluation, the prediction scores and the human annotations of the second dataset can be combined. In the parts where the human annotation identified a hateful instance, they overwrite the given predictions. Classes that had to be excluded due to definition mismatches can be annotated, but only with the predictions of the network. The predicted values are normalized to [0,1], while the human annotations remain binary.

8. Based on this mix of predicted and human-based annotations, the original network is fine-tuned again on the new dataset (see section 6.6). Extra measures to prevent overfitting can be implemented.

9. The dataset is evaluated again using the same binary hate/no-hate external dataset (see section 6.6).

10. Lastly, the two measurements of prediction quality on the external dataset are compared to validate the performance and provide insight into generalizability (see section 6.7).

4 Datasets

Two primary datasets with different annotations were selected for this research, along with two additional datasets: one for evaluation and one for balancing the two main datasets during training with non-hateful statements.

The first main dataset, provided by Vidgen et al. (2021), is a large, dynamically generated collection of 41,255 entries created over four rounds, with 54% of the entries being hateful. The dataset includes 11 English-language training datasets for hate and toxicity from hatespeechdata.com. Its hierarchical taxonomy, based on Robert C Nickerson and Muntermann (2013), classifies entries into hate and no-hate categories. The hate entries are further divided into five types (Derogation, Animosity, Threatening Language, Support for Hateful Entities, Dehumanization). Additionally, 29 identities as hate targets are annotated. The annotations were performed by 20 trained annotators.

The second main dataset compiled by Fanton et al. (2021) is also a dynamically generated human-in-the-loop dataset, containing 5,000 hateful statements. Created over two cycles with

human input in between, the initial dataset included 880 statements and was developed in collaboration with 20 experts from various NGOs. The annotations featured 10 labels ("DISABLED," "JEWS," "OVERWEIGHT," "LGBT+," "MUSLIM," "WOMEN," "PEOPLE OF COLOR," "ROMANI," "MIGRANTS," "OTHER"). Three trained students were involved in the annotation process.

The dataset from Fillies et al. (2023b) was selected for non-hateful statements, as only the hateful entries were selected from the two main datasets, and training a classifier solely on those would likely result in overfitting. This dataset, in English, includes annotated Discord messages collected between March 2021 and June 2022, comprising 88,395 chat messages. Around 6.42% of the messages were classified as hate speech.

The final support dataset, from Ljubešić et al. (2021), was chosen for validation and independent evaluation of the classifier's performance. It consists of YouTube comments collected between January and May 2020, with approximately 50% hate and 50% non-hateful examples.

5 General Taxonomy

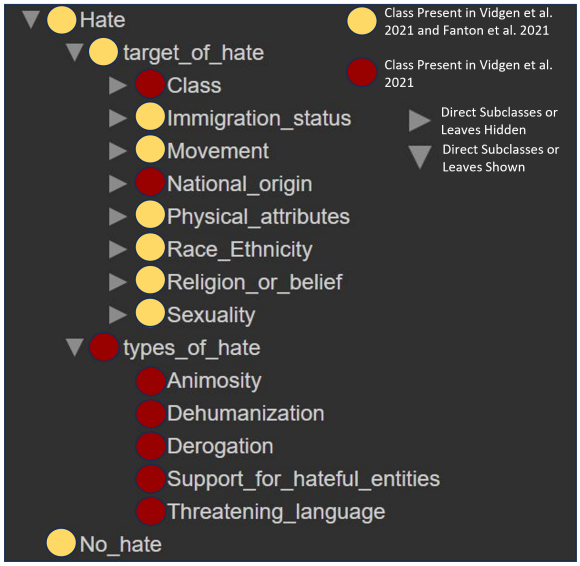

Figure 2: Overview General Taxonomy Level 1 - 3

This research explores merging multiple taxonomies into a central one to enable a single classifier to predict diverse definitions using differently annotated datasets. As a demonstration, two existing taxonomies were combined. The taxonomy was developed by a two-person team and is shown in Appendix A.1, with the first three levels in Figure 2. Shared classes and leaves (labels not further broken down) are highlighted in yellow, while those unique to Vidgen et al. (2021) are in red. Both taxonomies contributed different, identical, or new subclasses and leaves. The final taxonomy has five layers.

The taxonomy from Vidgen et al. (2021) formed the basis for the merge due to its thoroughness. It initially distinguishes between hate and non-hate statements.

Hate types from Vidgen et al. (2021) were grouped under the label "types_of_hate," which was absent in Fanton et al. (2021). Adjustments were made for hate targets, with seven out of 11 classes from Fanton et al. (2021) fitting directly into the new taxonomy. The remaining classes, like "Gender," "Intersectional," and "Disability," required modifications.

Due to Fanton et al. (2021) introducing the labels "Disabled" and "Overweight", a class regarding physical attributes was introduced, also containing the label "Gender," which then includes the class "Gender Minorities," unlike Vidgen et al. (2021) where it is independent. The last label from Vidgen et al. (2021), "Intersectional," was not included explicitly, as it is contained in the multi-label encodings (e.g., black women) that are represented in the taxonomy.

The classes ("Jews", "Muslim", "Women", "Romani", "Migrants") from Fanton et al. (2021) were already covered in the taxonomy. The label "People of color" from Fanton et al. (2021) was initially introduced as an independent label under the class "Physical_attributes/skin_color/" next to the labels "Black" and "White." However, the evaluation of the trained network's performance clearly showed this as a mistake, making it necessary to make "Black" a subclass of "People of Color."

The main challenge was the label "LGBT+" by Fanton et al. (2021) due to its covering of multiple aspects. It is first a political and social movement, standing for "lesbian, gay, bisexual, transgender, plus other sexual and gender identities," making it difficult to locate in the existing classes of gender and sexual orientation. The decision was made to include it in the taxonomy as a movement.

It is noteworthy that in the actual dataset annotations by Vidgen et al. (2021), labels appeared that were not represented in the provided taxonomy, such as "old.people," "russian," "lgbtq," "eastern.europe," and "non.white." These labels were included in the new general taxonomy with their own

classes. However, the label "other" from Fanton et al. (2021) had to be disregarded. The final taxonomy consists of 23 classes and 43 leaves, merging labels from both taxonomies directly or through abstraction.

6 Experimental Classifier

This section describes the creation of an experimental classifier. The classifier proves the validity of the concept as a proof-of-work. As detailed in the methodology section (3), the labels in the existing datasets from Vidgen et al. (2021) can be matched to the labels of the new taxonomy, creating a new annotation schema for the dataset. The annotated dataset is then used to fine-tune a pretrained language model to be a multi-label hate speech classifier. After this initial training, the classifier is used to reannotate the second dataset from Fanton et al. (2021), introducing the new annotation schema and providing insights into the created taxonomy, missing labels, and different underlying definitions of hate contained in the two datasets.

The predicted annotations can then be merged with the existing human annotations and used to fine-tune the network again. If the approach holds merit, the minimum requirement is that the hate speech prediction quality of the network increases on an independent test set after the training cycles. This section describes the steps of this process.

6.1 Encoding

The goal is to map the taxonomy into a network-readable format while preserving class structure information and enabling the annotation of multiple definitions within a unified schema. The proposed encoding uses a sparse binary vector, where each position corresponds to a class or leaf in the taxonomy. This allows the network to learn parent-child relationships while capturing varying degrees of hate within a single framework.

For example, in the schema "Target_of_hate / Physical_attributes / Skin_color / People_of_color / Black," a statement expressing hate toward Black people would be encoded as [1,1,1,1,1], while hate toward people of color would be [1,1,1,1,0]. This approach enables the network to recognize hierarchical relationships and adapt to different depths of hate speech definitions.

6.2 Evaluation Metrics

Two evaluation metrics were used: accuracy and F-1 scores. For a deeper understanding of the results, the distributions of predictions in regard to the human-annotated labels were evaluated in the four groups: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Accuracy is defined as the ratio of correct predictions to the number of total predictions. The F1-Score metric is beneficial in situations where datasets have imbalanced class distributions (Tsourakis, 2022), fitting the problem at hand. For the F-1 Score, a threshold of 0.5 was chosen.

6.3 Algorithm

As a base, the state-of-the-art model RoBERTa was chosen, first introduced by Liu et al. (2019b). It is a fine-tuned, improved version of the BERT model pretrained and introduced by Devlin et al. (2019). RoBERTa uses the same architecture as BERT but applies a different tokenizer and pretraining scheme. The research used the pretrained multi-label RoBERTa model for multi-label sequence classification provided through the platform HuggingFace.[2] In combination with the fitting tokenizer from "twitter-roberta-base-emotion".[3]. It is meant to be an example implementation to show merit.

6.4 Technical Setup

For training, Google Colaboratory (Colab) was used, providing a browser-based environment for writing and executing Python code in Jupyter notebooks. As noted by Kimm et al. (2021), Colab offers access to TPUs and GPUs without requiring additional configuration. For all training sessions, a cluster with Nvidia V100 GPUs, 12.7 GB System-RAM, 16 GB GPU-RAM, and 72.8 GB Storage was utilized. The first training cycle took 45 minutes, while the second cycle took 5 minutes. For both cycles, a fixed seed was used, with the evaluation step size set to 500, train and evaluation batch sizes set to 6, and the number of training epochs set to 4. Other hyperparameters followed the default recommendations from RoBERTa. To prevent overfitting during the second cycle, the dropout ratio for attention probabilities and the dropout probability for fully connected layers in embeddings, encoder, and pooler were both set to 0.5.

[2]https://huggingface.co/docs/transformers/model_doc/roberta#transformers. RobertaForSequenceClassification

[3]https://huggingface.co/cardiffnlp

6.5 Data preparation

As detailed in Section 6.1, both datasets were encoded using sparse one-hot encoding based on the taxonomy. They were cleaned of duplicates, missing data, and unusable annotations. Given the nature of BERT models, no additional text preprocessing was performed to preserve information. Since both datasets lacked non-hateful language, 30% non-hateful statements from Fillies et al. (2023b) were randomly added. Considering that only 6% of the 88,000 messages in Fillies et al. (2023b) contain hate, the risk of including complex cases like counter-hate speech was minimal. These non-hateful examples were also one-hot encoded. A 10% holdout set was reserved for evaluation, and both datasets were randomized.

After cleaning and adding 30% non-hate speech statements, the dataset from Vidgen et al. (2021) contained 18,380 instances, while the dataset based on Fanton et al. (2021) had 4,767 instances.

The annotation of the Fanton et al. (2021) dataset combined human annotations from Fanton et al. (2021) with predictions from the first training cycle. When the network failed to predict a label but an annotator identified it, the human annotation took precedence. This approach is justified, as human annotations rely on inter-annotator agreement, reducing the likelihood of false positives, since multiple annotators would need to select the same incorrect label. When no human labels were available or the human annotation didn't match the network's prediction, the network's predictions were used. This was necessary because certain labels were not annotated in the second dataset, and false negatives by annotators were more likely, given that inter-annotator agreement was reduced to binary decisions. For example, the network might predict a low likelihood of racism in a statement (e.g., a score of 0.2 on a scale from -1 to 1). However, human annotation, based on a binary majority agreement among three annotators (two say no racism detected, but one identifies racism), could be flawed. In such cases, the network's prediction is considered a more accurate reflection of reality than the potentially flawed binary annotation.

6.6 Results

The prediction results from the three fine-tuning experiments and their evaluation on the independent evaluation test set (ETS) are shown in Table 1. The details of these results are discussed individually in section 6.7.

Table 1: All Training and Evaluation Test Set Results

Cycle	Dataset	F1-Score	Accuracy
Cycle-1	Vidgen	0.89	0.46
Cycle-1	ETS	0.73	-
Cycle-1-A	Vidgen	0.89	0.55
Cycle-1-A	ETS	0.73	-
Cycle-2	Fanton	0.91	0.74
Cycle-2	ETS	0.84	-

Table 2: Display of selected classes from the class wise prediction's evaluation of RoBERTa-Cycle-1 on the dataset by (Fanton et al., 2021)

Class/Leaves	F1-Score	Instances
Hate	1.00	3539
Target_of_hate	0.99	3539
Movement	0.00	465
LGBTQ+	0.00	465
Physical_attri	0.90	1036
Skin_color	0.93	301
Black	0.00	0
Non_white	0.03	301
Religion/belief	0.99	1401
Jews	0.99	418
Muslims	0.98	983
Sexuality	0.00	0
Bisexual	0.00	0
Gay	0.00	0
Types_of_hate	0.00	0
Weighted avg	0.89	15017

6.6.1 RoBERTa-Cycle-1

In the first stage, the classifier (RoBERTa-Cycle-1) was trained on the dataset from Vidgen et al. (2021) and evaluated on the evaluation dataset from Ljubešić et al. (2021).

This training and evaluation were followed by an analysis of the classifier's predictions at the class level for the dataset from Fanton et al. (2021) (see Table 2). For each class, results were assessed, and performance drops, such as in the cases of 'Non_white' and 'LGBTQ+', were identified. Incorrectly associated labels were pinpointed (see Table 3 and 4). For instance, many statements labeled 'LGBTQ+' were misclassified under the "Sexuality" label. Table 3 shows the percentages of other classes predicted for the "LGBTQ+" label, while Table 4 shows the misclassification for "Non_white". The percentages do not add up to

Table 3: Display of selected classes where the class "LGBTQ+" gets miss labeled to. Using the RoBERTa-Cycle-1 model on the dataset by Fanton et al. (2021)

Class	Percentage
Physical_attri.	0.308
Gender	0.295
Gender_min.	0.189
Trans	0.166
Women	0.037
Sexuality	0.850
Gay	0.819

Table 4: Display of selected classes where "Non_white" is mislabeled, using the RoBERTa-Cycle-1 model on the dataset by (Fanton et al., 2021).

Class	Percentage
Black	0.882
Race_Ethnicity	0.078

1, as this is a multi-label prediction with binary annotations.

These misclassifications highlight the need for adjustments in the taxonomy, as "LGBTQ+" and "Non_white" are not correctly represented. This led to the need to relabel and retrain the model, resulting in RoBERTa-Cycle-1-A.

6.6.2 RoBERTa-Cycle-1-A

In the following, the model RoBERTa-Cycle-1-A and its performance on the Evaluation Test were established, see Table 1. It can be observed that the F-1 score remains stable while the accuracy increases significantly after adjusting the taxonomy. All prediction results for all classes of the datasets can be found on GitHub[4]. Table 5 displays a selection of classes important for evaluating the adjustment of the taxonomy in the previous step.

After the training of RoBERTa-Cycle-1-A, the same in-depth evaluation of the classifier's predictions on a class level for the dataset from Fanton et al. (2021) was performed, see GitHub[5].This time, no outlier class, in terms of prediction performance, was identified, indicating that there is no further need for adjustment.

[4]https://github.com/fillies/HateSpeechCrossTaxonomyDatasetIntegration

[5]https://github.com/fillies/HateSpeechCrossTaxonomyDatasetIntegration

Table 5: Display of selected classes from the class wise predictions evaluation of RoBERTa- Cycle-1-A on the dataset by (Fanton et al., 2021)

Class/Leaves	F1-Score	Instances
Hate	1.00	3539
Target_of_hate	0.99	3539
Skin_color	0.94	301
Non_white	0.94	301
Black	0.00	0
Weighted avg	0.91	-

6.6.3 RoBERTa-Cycle-2

Based on RoBERTa-Cycle-1-A and the merged machine and human annotations of the Fanton et al. (2021) dataset, the model RoBERTa-Cycle-2 was trained and evaluated on the Evaluation Test Set, see Table 1. A relevant increase in F1-Score (from 0.73 to 0.84) on the ETS can be observed, accompanied by a general increase in prediction quality on the new dataset (to a new F1-Score of 0.91 and an accuracy of 0.74).

Different from RoBERTa-Cycle-1 and similar to RoBERTa-Cycle-1-A, the evaluation of each annotated class and its prediction performance, see Table 6, did not produce noteworthy outliers in regard to underperformance. Therefore, no further adjustment of the taxonomy is necessary. All prediction results for all classes across all datasets can be found on GitHub[6].

6.7 Discussion of Results

6.7.1 RoBERTa-Cycle-1

After the first training cycle on the dataset from Vidgen et al. (2021), the results in table 1, particularly the F1-Score, show strong performance for the RoBERTa-Cycle-1 classifier. The notable difference between F1-Score and Accuracy highlights the class imbalance, which corresponds with the sparse input vectors and unbalanced class distributions in the dataset. The F1-Score of 0.73 on the Evaluation Test Set further confirms that the classifier successfully learned and generalized the key aspects of hate speech.

The predictions from RoBERTa-Cycle-1 on the Fanton et al. (2021) dataset (see Table 2) show that the model excels at identifying higher levels of abstraction, especially in binary hate speech classification, but struggles with more specific categories.

[6]https://github.com/fillies/HateSpeechCrossTaxonomyDatasetIntegration

Table 6: Display of selected classes from the class wise predictions evaluation of RoBERTa-Cycle-2 on the dataset from Vidgen et al. (2021)

Class/Leaves	F1-Score	Instances
Hate	1.00	14900
Target_of_hate	1.00	14780
Movement	0.00	0
LGBTQ+	0.00	0
Physical_attributes	0.93	7541
Skin_color	0.88	2918
Black	0.86	2553
Non_white	0.89	2918
Religion_or_belief	0.86	2529
Jews	0.87	1293
Muslims	0.84	1267
Sexuality	0.89	1552
Bisexual	0.00	110
Gay	0.87	1487
Types_of_hate	1.00	14900
Weighted avg	0.82	-

Three issues are observed. First, annotations, such as "types_of_hate," are missing from the Fanton et al. (2021) annotations.

Second, while the network performs well in predicting the "skin_color" class, it mislabels many "non_white" statements as "black," indicating a taxonomy error (see Table 4). The error rate of around 11% across other classes is acceptable given the network's overall performance. Lastly, the network significantly underperforms on the "Movement" class and the "LGBTQ+" leaf, with misclassifications spread across multiple leaves in different classes (see Table 3), suggesting a mismatch in definitions. The issue of mismatched definitions is a clear limitation at this stage. For cases like "black" and "non_white," taxonomy adjustments—such as making "non_white" the parent class of "black"—can help address misclassifications within leaves or subclasses. However, deeper issues, like the "LGBTQ+" misclassifications, may require more advanced solutions, potentially utilizing ontology matching techniques in the future.

6.7.2 RoBERTa-Cycle-1-A

After retraining the classifier with the new encoded filtered input, Table 1 shows improved accuracy for RoBERTa-Cycle-1-A and resolves the taxonomy issue for "black" and "non_white" classes (see Table 5). This performance increase is linked to the label adjustment based on the revised taxonomy. The network's prior learning that "black" is a leaf of "non_white" highlights the value of encoding semantic relationships into labels, enhancing label comparability and generalizability in future iterations.

6.7.3 RoBERTa-Cycle-2

RoBERTa-Cycle-2's class-wise performance on the dataset from Vidgen et al. (2021) (see Table 6) shows that, despite retraining, it preserves the original class definitions (e.g., "types_of_hate") while improving its general understanding of hate speech, as evidenced by the increase in prediction quality on the Evaluation Test Set from 0.73 to 0.84.

Although there is a slight decrease in the weighted average prediction quality from 0.89 to 0.82 on the Vidgen et al. (2021) dataset, this is reasonable given the complete fine-tuning. The model adapts well, correctly covering both new and old concepts, demonstrating that careful design and fine-tuning allow it to retain learned patterns while adapting to new definitions.

7 Conclusion and Outlook

The results of this research demonstrate the feasibility of combining different hate speech taxonomies into a single, general taxonomy, which can be used to train a classifier capable of predicting a broader range of hate speech definitions. This approach reduces the need for multiple niche models, minimizing computational resources, and allows for model training without sharing sensitive data, thus addressing privacy concerns. The semantic relationships encoded in the labels also enhance generalizability for further training, aligning with current research in federated learning and continuous learning for hate speech detection.

By iteratively fine-tuning a pre-trained multi-label classifier on two distinct datasets, the research shows that a general taxonomy can improve hate speech detection, leading to higher performance in classifying general hate speech, as demonstrated on an independent evaluation test set. This work serves as proof that a general taxonomy can be used in multi-label hate speech classification, integrating diverse datasets and definitions of hate speech. It also suggests that, in the future, only trained networks need to be exchanged, not the sensitive datasets, advancing federated hate speech detection.

Looking ahead, further research is needed to explore automatic matching of taxonomies on both

logical and semantic levels, including detecting mismatches based on definitions. Validation with a broader variety of hate taxonomies, and possibly the creation of a hate speech ontology, is essential. Additionally, encoding structural knowledge through ontologies holds significant potential. Further work is needed on bias mitigation and quality assurance in the context of hate speech detection.

Limitations

The work has to address the following limitations. Firstly, it does not serve as a general proof that all datasets and all taxonomies can be combined into one. As seen in the work already, certain subparts of the two choose example taxonomies could not be merged. The problems seen here are similar to the problems arising and handled within the ontology matching community (Shvaiko and Euzenat, 2013), the found solutions from that field will greatly contribute to future development of the approach. Furthermore, a significant challenge is that at least the first round of training is done with possibly mislabeled data, which could lead to underperformance in the field. Similarly, the usage of algorithmically created annotations may propagate biases and underperformance, potentially even enhancing them. Lastly, the proposed iterative retraining could lead to the loss of the originally trained definitions of hate and functionality, if no countermeasures, such as more advanced subclass test sets and overfitting prevention, are conducted.

Ethical Considerations

Even though machine learning based applications to detect hate speech automatically online are not the solution to hate online, they are a fundamental tool in the process of combating online hate speech. this research advocated for a contextual aware human-in-the-loop strategy to counter online hate speech. The research is in the interest of society, and the public good is a central concern. The algorithmic detection of hate speech is necessary to provide a harm-free space, especially for demographic groups with special needs for protection, such as adolescents. The research is advancing the field in a more open but data-secure direction. While more diverse understandings of what constitutes hate speech is usable, the potential limitations are stated in section 7.

References

Valentina Bartalesi, Rachele Sprugnoli, Valentina Bartalesi Lenzi, and Giovanni Moretti. 2006. Cat: the celct annotation tool creep (cyberbullying effects prevention) view project it-timebank view project cat: the celct annotation tool.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stacy Jo Dixon. 2023. Facebook hate speech removal per quarter 2023.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. 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 3226–3240, Online. Association for Computational Linguistics.

Jan Fillies, Michael Peter Hoffmann, and Adrian Paschke. 2023a. Multilingual hate speech detection: Comparison of transfer learning methods to classify german, italian, and spanish posts. In *2023 IEEE International Conference on Big Data (BigData)*, pages 5503–5511. IEEE.

Jan Fillies and Adrian Paschke. 2024. Simple llm based approach to counter algospeak. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 136–145.

Jan Fillies, Silvio Peikert, and Adrian Paschke. 2023b. Hateful messages: A conversational data set of hate speech produced by adolescents on discord. *Preprint*, arXiv:2309.01413.

Jan Fillies, Esther Theisen, Michael Hoffmann, Robert Jung, Elena Jung, Nele Fischer, and Adrian Paschke. 2025. A novel german tiktok hate speech dataset: far-right comments against politicians, women, and others. *Discover Data*, 3(1):4.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Jay Gala, Deep Gandhi, Jash Mehta, and Zeerak Talat. 2023. A federated approach for hate speech detection. *arXiv preprint arXiv:2302.09243*.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *Social Informatics*, pages 49–66, Cham. Springer International Publishing.

Lennart Justen, Kilian Müller, Marco Niemann, and Jörg Becker. 2022. No time like the present: Effects of language change on automated comment moderation. In *2022 IEEE 24th Conference on Business Informatics (CBI)*, volume 01, pages 40–49.

Simon Kemp. 2023. Digital 2023 april global statshot report - datareportal – global digital insights.

Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. Conprompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10964–10980.

Haklin Kimm, Incheon Paik, and Hanke Kimm. 2021. Performance comparision of tpu, gpu, cpu on google colaboratory over distributed deep learning. *2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, pages 312–319.

Hannah Rose Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in nlp research. *arXiv preprint arXiv:2204.14256*.

Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Olga Dontcheva-Navrátilová, Chaya Liebeskind, Giedrė Valūnaitė Oleškevičienė, Slavko Žitnik, Marcin Trojszczak, Renata Povolná, Linas Selmistraitis, Andrius Utka, et al. 2023. Llod schema for simplified offensive language taxonomy in multilingual detection and applications. *Lodz papers in pragmatics*, 19(2):301–324.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Nikola Ljubešić, Igor Mozetič, Matteo Cinelli, and Petra Kralj Novak. 2021. English YouTube hate speech corpus. Slovenian language resource repository CLARIN.SI.

Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F. Johnson. 2023. Offline events and online hate. *PLOS ONE*, 18(1):1–14.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PLOS ONE*, 15(8):1–26.

Isar Nejadgholi and Svetlana Kiritchenko. 2020. On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online. Association for Computational Linguistics.

Ali Omrani, Alireza S Ziabari, Preni Golazizian, Jeffery Sorensen, and Morteza Dehghani. 2023. Towards a unified framework for adaptable problematic content detection via continual learning. *arXiv preprint arXiv:2309.16905*.

Flor Miriam Plaza-del Arco, Debora Nozza, Dirk Hovy, et al. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics.

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*, 55:477–523.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Upkar Varshney Robert C Nickerson and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Sanjana Sharma, Saksham Agrawal, and Manish Shrivastava. 2018. Degree based classification of harmful speech using Twitter data. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 106–112, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pavel Shvaiko and Jérôme Euzenat. 2013. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):158–176.

Adriana Stephan. 2020. Comparing platform hate speech policies: Reddit's inevitable evolution. A program of the Cyber Policy Center, a joint initiative of the Freeman Spogli Institute for International Studies and Stanford Law School.

Marco Antonio Stranisci, Simona Frenda, Mirko Lai, Oscar Araque, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, and Viviana Patti. 2022. O-dang! the ontology of dangerous speech messages. In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, pages 2–8, Marseille, France. European Language Resources Association.

Nikos Tsourakis. 2022. *Machine Learning Techniques for Text: Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation*. Packt Publishing.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2019. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018*, pages 1 – 10. Austrian Academy of Sciences, Vienna, Austria.

Marcos Zampieri, Damith Premasiri, and Tharindu Ranasinghe. 2024. A federated learning approach to privacy preserving offensive language identification. *arXiv preprint arXiv:2404.11470*.

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 53–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Appendix

A.1 The General Taxonomy

The general taxonomy has on level 0 the classes Hate and No-hate. On level 1 it is further broken down into Target_of_hate and Types_of_hate.

1. No-hate
2. Hate
   (a) Target_of_hate
   (b) Types_of_hate

Target_of_hate is further broken down into:

1. Class
   (a) Working_class
2. Immigration_status
   (a) Asylum_seeker
   (b) Foreigner
   (c) Immigrants
   (d) Refugee
3. Movement
   (a) LGBTQ+
4. National_origin
   (a) China
   (b) Korea
   (c) Pakistan
   (d) Other_N
   (e) Poland
   (f) Russian
5. Physical_attributes
   (a) Age
      i. Old
      ii. Young
   (b) Disability
   (c) Gender
      i. Gender_minorities
         A. Trans

    ii. Man
    iii. Women
  (d) Overweight
  (e) Skin_color
    i. Black
    ii. Non_white
    iii. White

6. Race_Ethnicity
  (a) Arabs
  (b) Asia
    i. East_A
    ii. South
    iii. South_east
  (c) Black_people
  (d) Europe
    i. East_E
  (e) Hispanic
  (f) Indigenous
    i. Aboriginal_people
  (g) Minority_groups
  (h) Mixed_race
  (i) People_from_Africa
  (j) Travelers
    i. Roma

7. Religion_or_belief
  (a) Hindus
  (b) Jews
  (c) Muslims
  (d) Other_R

8. Sexuality
  (a) Sexuality
  (b) Bisexual
  (c) Gay
  (d) Lesbian

Types_of_hate is further broken down into:

1. Animosity
2. Dehumanization
3. Derogation
4. Support_for_hateful_entities
5. Threatening_language