Analyzing Offensive Language Dataset Insights from Training Dynamics and Human Agreement Level

Do-Kyung Kim[1], Hyeseon Ahn[1], Youngwook Kim[2] and Yo-Sub Han[1]*
[1]Department of Computer Science, Yonsei University, Seoul, South Korea
[2] KT, Seoul, South Korea
{kdky95, hsan, emmous}@yonsei.ac.kr
young-wook.kim@kt.com

Abstract

Implicit hate speech detection is challenging due to its subjectivity and context dependence, with existing models often struggling in out-of-domain scenarios. We propose CONELA, a novel data refinement strategy that enhances model performance and generalization by integrating human annotation agreement with model training dynamics. By removing both easy and hard instances from the model's perspective, while also considering whether humans agree or disagree and retaining ambiguous cases crucial for out-of-distribution generalization, CONELA consistently improves performance across multiple datasets and models. We also observe significant improvements in F1 scores and cross-domain generalization with the use of our CONELA strategy. Addressing data scarcity in smaller datasets, we introduce a weighted loss function and an ensemble strategy incorporating disagreement maximization, effectively balancing learning from limited data. Our findings demonstrate that refining datasets by integrating both model and human perspectives significantly enhances the effectiveness and generalization of implicit hate speech detection models. This approach lays a strong foundation for future research on dataset refinement and model robustness. [1]

1 Introduction

Implicit hate speech is subtle and often hidden in seemingly harmless language, requiring nuanced interpretation to detect. This makes implicit hate speech detection a challenging problem in natural language processing due to its inherent subtlety and context-dependent nature. Unlike explicit hate speech, which can be easily identified through apparent offensive language, implicit hate speech often relies on indirect expressions, metaphors, and coded language, making it more difficult to detect accurately (Wiegand et al., 2021; ElSherief et al., 2021). The ambiguity and subjectivity associated with implicit hate speech further complicate the task, leading to inconsistencies in human annotations and reduced performance in detection models (MacAvaney et al., 2019).

Although there has been substantial progress in detecting explicit hate speech, language models still struggle to generalize themselves when applied to implicit cases, especially in out-of-domain scenarios (Bourgeade et al., 2023). Many existing models exhibit fair results when tested on in-domain datasets; however, they often fail to maintain similar performance on external datasets. This lack of generalization is particularly concerning in implicit hate speech detection, where context and subtle cues can vary greatly between datasets, making it critical for models to effectively handle unseen data (Mathew et al., 2021).

These challenges motive us to propose a better data selection and refinement strategy to enhance the generalizability of implicit hate speech detection models. Swayamdipta et al. (2020) focused on improving the data quality to enhance model performance based on training dynamics—the behavior of a model throughout the training process. Training dynamics are determined by model confidence and variability during the training step, and are used to characterize the learning difficulty of data instances. For our task, we utilize training dynamics to split the training dataset into three parts—Easy-to-Learn (EtL), Ambiguous-to-Learn (AtL), and Hard-to-Learn (HtL). Then, we quantify the agreement level among annotators for each data sample, and separate the instances into two cases—consensual and non-consensual cases. Consensual instances are the ones that annotators largely agree on the data label, which implies a clear judgment. Non-consensual instances are the ones that annotators give differing labels, which implies ambiguity

*Corresponding Author.
[1]Our code is available at https://github.com/kdkcode/CONELA.

Proceedings of the 31st International Conference on Computational Linguistics, pages 9780–9792
January 19–24, 2025.

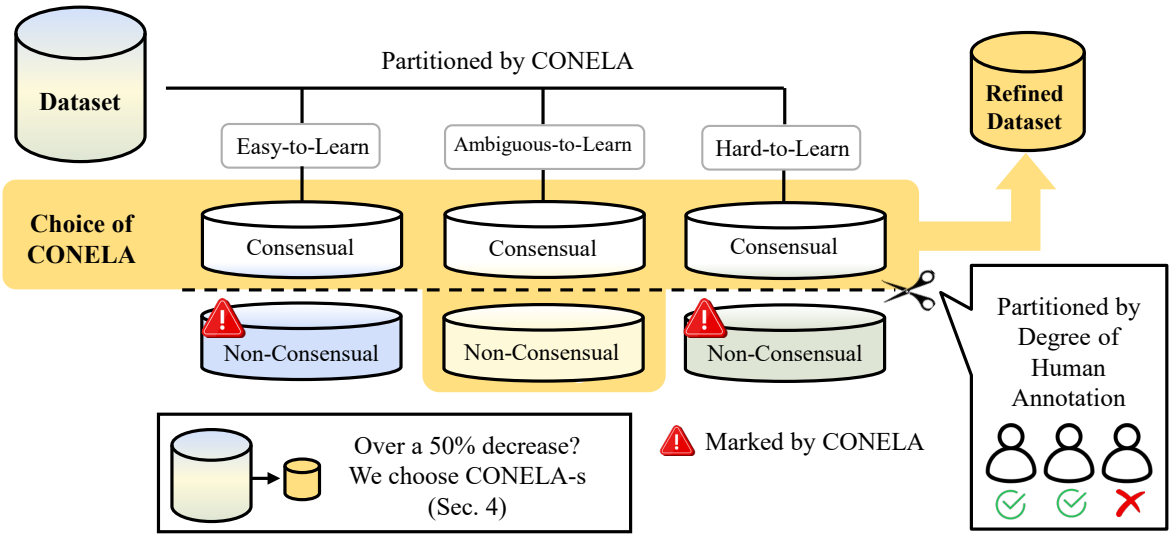


Figure 1: An overview of our data selection strategy. CONELA categorizes data into Easy, Hard, and Ambiguous-to-Learn sets using model dynamics and human agreement. It retains consensus instances, especially ambiguous ones, while discarding disagreements in Easy and Hard sets. This refinement enhances data quality for implicit hate speech detection models.

or subjectivity of the data sample. We discuss the concept of consensuality in detail in Section 3.2.

Figure 1 illustrates an overview of our approach. By removing instances from Easy-to-Learn (EtL) and Hard-to-Learn (HtL) categories where annotators disagree, we retain ambiguous cases that are crucial for out-of-distribution generalization. We call the approach **CONELA**: Consensual elimination Of Non-Consensual EtL and HtL Annotations. In addition, we adopt a weighted loss function that incorporates the existing adversarial disagreement maximization (ADM) method (ADM, Lee et al., 2023) alongside an ensemble strategy to cope with the case when the training data size is relatively small. This combination helps mitigate the negative impact of removing too much data from smaller datasets while still leveraging the benefits of excluding non-consensual instances. The experimental results show that CONELA consistently improves model performance across multiple datasets.

Our contributions are threefold:

1. We introduce CONELA, a refined data selection methodology that incorporates training dynamics and human annotation agreements to improve model robustness.

2. We conduct extensive experiments across various model and dataset combinations to determine the most effective approach, providing robust insights into the optimal methodology.

3. We propose a weighted loss function and an ensemble strategy to address data scarcity issues and maintain balanced learning when working with smaller datasets.

2 Related Works

2.1 Hate Speech Detection

Hate speech detection has evolved significantly, from early lexicon-based methods (Ding et al., 2008; Bonta et al., 2019) to context-aware approaches (Gao and Huang, 2017; Pérez et al., 2023). Recent focus has shifted towards improving generalization (Kim et al., 2022; Hong and Gauch, 2023; Ahn et al., 2024) through data augmentation and reannotation techniques (Zhou et al., 2021; Tal and Vilenchik, 2022). Advancements in large language models and transfer learning have led to more comprehensive datasets like HateXplain (Mathew et al., 2021) and ToxiGen (Hartvigsen et al., 2022). Implicit hate speech detection remains challenging, driving the development of advanced techniques like DeepHate (Cao et al., 2020) and benchmarks designed to capture subtle forms of hatred (Sap et al., 2020; ElSherief et al., 2021). While these efforts have made meaningful progress, we believe that tackling the difficulties of implicit hate speech detection requires a closer examination of both model performance and human interpretation. This perspective aligns with recent studies exploring implicitly abusive language (MacAvaney et al., 2019;

Vidgen et al., 2021; Caselli et al., 2020; Wiegand et al., 2021). Our work introduces a new approach that incorporates model training dynamics and human annotation agreement to improve the generalizability of implicit hate speech detection systems.

2.2 Data Maps

Data maps are a visualization tool that leverages training dynamics to characterize different regions of a dataset (Swayamdipta et al., 2020; Zhang and Plank, 2021). Swayamdipta et al. (2020) considered the following three key regions in a map:

- Easy-to-Learn (EtL): Instances quickly learned by the model
- Ambiguous-to-Learn (AtL): Instances contributing to out-of-distribution generalization
- Hard-to-Learn (HtL): Challenging instances that are often mislabeled

These regions are determined by confidence and variability. We utilize these regions by dividing each region further into two sub-regions based on the degree of human annotation agreements. We establish six distinct zones that enable to have a more nuanced analysis of how different types of instances impact the model learning and generalization.

By combining model-based measures (from training dynamics) with human-based measures (annotation agreement), we aim to provide a more comprehensive understanding of dataset characteristics and their influence on model behavior. This granular segmentation may offer new insights into dataset quality, model learning patterns, and strategies for dataset curation and model training (Toneva et al., 2019; Pleiss et al., 2020; Paul et al., 2021; Kwon and Zou, 2023).

We observe that the distribution of data instances varies significantly between a general dataset (SNLI) and an implicit hate speech dataset as depicted in Figure 2.

2.3 Subjectivity and Disagreement in Data Annotations

Recent research focuses on the importance of embracing subjectivity and disagreement in data annotations for complex tasks including hate speech detection. Leonardelli et al. (2023) and Uma et al. (2021) examined a softer approach to subjective labels instead of the traditional pursuit of perfect inter-annotator agreement. Pavlick and Kwiatkowski (2019) and Basile et al. (2021) argued that disagreement often reflects genuine linguistic ambiguity and should be preserved. Sap et al. (2020) highlighted how annotator beliefs can bias toxic language detection, while Röttger et al. (2022) proposed contrasting annotation paradigms for subjective NLP tasks. Previous studies (Solorio and Liu, 2008; Zhou and Li, 2010; Leonardelli et al., 2021) showed that training on multiple annotations even with disagreement can improve the model performance and generalization. Meissner et al. (2021) and Hartvigsen et al. (2022) demonstrated how to use disagreement information to improve model performance in subjective tasks. We adopt the approach that embraces annotation subjectivity using

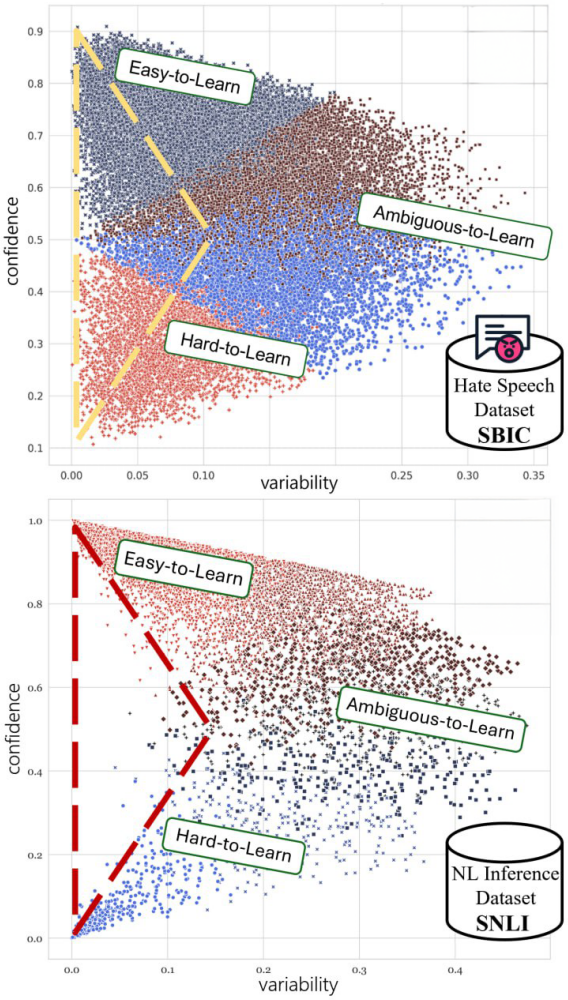


Figure 2: The data map of the SBIC dataset (top) and the SNLI dataset (bottom). The triangular area is expected to remain empty because the dataset is divided into three regions based on confidence and variability. If the triangular area is populated, then it indicates ambiguity in the boundaries between confidence and variability.

data maps (Swayamdipta et al., 2020) to classify instances based on the model confidence and the human annotator agreement.

3 CONELA: Our Data Selection Strategy

We propose CONELA to consider both consensual and non-consensual aspects of human annotations. As depicted in Figure 1, CONELA first divides the dataset into three regions–EtL, AtL, and HtL–based on training dynamics. Then, according to the agreement level among annotators, CONELA further divides each region into two subregions—consensual and non-consensual areas—and gives rise to six regions in the end. We explain the training dynamics in Section 3.1 and how the agreement level is quantified in Section 3.2.

3.1 Training Dynamics Analysis

Training dynamics reveal learning characteristics that enable the division of instances into Easy-to-Learn (EtL), Ambiguous-to-Learn (AtL), and Hard-to-Learn (HtL) categories. These boundaries are defined using confidence C in Equation (1) and variability V in Equation (2):

$$C = \frac{1}{N}\sum_{i=1}^{N} P(y_i \mid x_i), \tag{1}$$

$$V = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(P(y_i \mid x_i) - C\right)^2}, \tag{2}$$

where $P(y_i|x_i)$ is the model's probability for the correct label y_i given i^{th} data sample x_i over N epochs. These metrics help differentiate EtL, AtL, and HtL data points, improving dataset integrity and model performance. Figure 2 illustrates these categories on our constructed data map.

Song et al. (2023) introduced a novel approach to identify noisy labels using training dynamics. Zhou et al. (2022) proposed a debiased training method based on data map analysis, and Min et al. (2022) presented a comprehensive study on the relationship between training dynamics and model generalization. In these trends, our research also leverages training dynamics for data analysis, providing insights into model behavior and dataset characteristics.

3.2 Consensual vs. Non-Consensual

We define the degree of human agreements to be

$$\text{Degree of human agreement} = \begin{cases} \textit{Consensual}, & \text{if all agreed;} \\ \textit{Non-Consensual}, & \text{otherwise.} \end{cases}$$

In other words, among three possible labels (offensive, non-hate and ambiguous). being consensual means that all annotators give the same label and being non-consensual means that the corresponding instance reflects subjectivity.

Our approach focuses on selectively removing data that introduces noise due to the lack of consensus between models and humans. Table 1 provides a comprehensive overview of all the categories used in this study, including their descriptions and the relationship between model and human perceptions. Specifically, we remove EtL-Non-Consensual and HtL-Non-Consensual instances, based on their potential to introduce misleading patterns during training.

- **EtL Non-Consensual**: These examples are learned quickly by the model but fail to align with human understanding, likely because they involve subtle nuances, such as implicit hate speech or biases. Excluding them helps prevent the model from becoming overly confident in misleading patterns.
- **HtL Non-Consensual**: These examples are difficult for both the model and human annotators to classify, often due to mislabeling or extreme complexity. Removing them reduces confusion during training, allowing the model to focus on more reliable instances.

We keep ambiguous instances in which both the model and humans show variability in classification. These instances are crucial for improving out-of-distribution generalization as demonstrated by works like Swayamdipta et al. (2020) and Kocon et al. (2023).

Categories	Description
EtL Consensual	Easy for both models and humans.
EtL Non-Consensual	Easy for models but difficult for humans.
AtL Consensual	Unclear for models but easy for humans.
AtL Non-Consensual	Unclear for both models and humans.
HtL Consensual	Difficult for models but easy for humans.
HtL Non-Consensual	Difficult for both models and humans.

Table 1: Description of data categories used in the study.

Dataset	Train	Refined Train (*ours*)	Test
SBIC	35,424	28,322	4,691
OLID	19,826	18,401	2,479
ETHOS	798	568	100
DYNAHATE	-	-	4,120
ToxiGen	-	-	8,960

Table 2: SBIC, OLID, ETHOS (with human agreement levels) for training; ToxiGen and DYNAHATE (without human agreement data) for evaluation only.

4 Addressing Data Scarcity

Addressing data scarcity in implicit hate speech detection is particularly challenging for smaller datasets (Rahman et al., 2021; Pal et al., 2022). Inspired by the adversarial disagreement maximization loss (Chen et al., 2024), we adopt and modify it in our approach and combine an ensemble of models to mitigate the issue.

Our method aims to simultaneously maximize model accuracy while increasing disagreement among ensemble models, thereby improving generalization across limited data. This approach is particularly effective for small datasets where overfitting is a significant risk (Li et al., 2021; Johnson and Khoshgoftaar, 2019). Our loss function consists of two components: the traditional cross-entropy loss and a disagreement loss. This combination ensures that different models within the ensemble make diverse predictions (Lee et al., 2023; Goodfellow et al., 2014). The overall loss is formulated as:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{Disagreement}}, \quad (3)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for the primary classification task. The disagreement loss is defined as:

$$\mathcal{L}_{\text{Dis}} = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{k=1}^{C}\text{softmax}(y_i^k)\cdot\text{softmax}(y_i'^k), \quad (4)$$

where N is the number of models in the ensemble, C is the number of classes, and y_i and y_i' represent the predictions of different models (Lakshminarayanan et al., 2017; Zhang et al., 2018).

Ensemble Strategy We incorporate an ensemble strategy (Xu et al., 2022; Dietterich, 2000) to improve the performance on small datasets. We generate predictions from two models: one trained on EtL non-consensual and HtL non-consensual data, and another on EtL consensual, AtL, and HtL consensual data. We then apply a majority voting mechanism defined as $\hat{y} = \text{mode}\left(\{\hat{y}_1, \hat{y}_2\}\right)$, where $\hat{y}_1$ and $\hat{y}_2$ represent the predictions from each model, and the final prediction $\hat{y}$ is determined by the majority vote.

5 Experimental Setup

We evaluate CONELA on five implicit hate speech datasets and use three of them for training. Our experimental results are averaged over five runs to ensure reliability and show significant performance improvements across various models.

For training, we use three implicit hate speech datasets together with the degree of human agreements—the Social Bias Inference Corpus (SBIC, Sap et al., 2020), Offensive Language Identification Dataset (OLID, Zampieri et al., 2019), and Online Hate Speech Detection Dataset (ETHOS, Mollas et al., 2022). Detailed information about the datasets can be found in Appendix A.

Then we evaluate both traditional transformer-based models and widely used large language models. We employ BERT (Devlin et al., 2019), HateBERT (Caselli et al., 2020), and RoBERTa (Liu et al., 2019) for transformer-based LMs, and LLaMA-3.1-8B[2], GPT-3.5-turbo[3], GPT-4o[4] and GPT-4o-mini[5] for LLMs. For these LLMs, we employ zero-shot prompting to adapt them to our hate speech detection task without fine-tuning. The prompt we used is provided in Appendix D.

6 Experimental Results and Discussion

Our experimental results demonstrate the effectiveness of CONELA in improving the performance of implicit hate speech detection models. In the following, we provide a detailed analysis of our findings focusing on the impact of removing EtL-non-consensual and HtL-non-consensual instances from the training data.

6.1 Performance Comparison

Table 3 presents the F1 scores of models trained on different datasets, comparing the baseline approach

[2] https://www.llama.com/

[3] https://platform.openai.com/docs/models/gpt-3-5-turbo

[4] https://openai.com/index/hello-gpt-4o/

[5] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Test	Train					
	SBIC		OLID		ETHOS	
	baseline	*Ours*	baseline	*Ours*	baseline	*Ours*
SBIC	85.52 ± 0.4	**87.56** ± 0.3	58.97 ± 1.8	**61.49** ± 2.2	58.93 ± 3.9	**66.59** ± 1.1
OLID	83.64 ± 0.9	**90.35** ± 1.2	97.21 ± 0.1	**97.28** ± 0.1	59.51 ± 6.2	**72.39** ± 2.5
ETHOS	72.47 ± 1.8	**76.59** ± 1.3	60.07 ± 1.3	**61.06** ± 2.3	68.56 ± 4.7	**71.99** ± 7.2
DYNAHATE	72.21 ± 0.3	**73.47** ± 0.3	53.51 ± 3.0	**55.57** ± 2.4	58.47 ± 3.3	**61.31** ± 4.3
ToxiGen	61.67 ± 0.7	**64.08** ± 1.0	38.82 ± 3.2	**39.63** ± 2.6	40.06 ± 4.2	**48.22** ± 2.5

Table 3: F1 Score performance comparison of BERT-uncased trained on SBIC, OLID, ETHOS dataset across different datasets and conditions. Note that both DynaHate and ToxiGen datasets below the dashed line are not used for training.

with our CONELA strategy. The results show several key insights regarding both in-domain (ID) and out-of-domain (OOD) performance.

For ID performance, CONELA yields marginal improvements in F1 scores across all datasets. This shows that the removal of non-consensual instances does not negatively impact the model's ability to classify data within the same domain. When evaluating OOD performance, CONELA significantly enhances the results particularly for certain dataset combinations. A detailed study of these dataset combinations is reported in Section 7. Models trained on SBIC exhibit the highest improvement with a maximum gain of 6.71%p in F1 score for test on ETHOS. OLID-trained models achieve up to a 2.52%p increase while models trained on ETHOS show the most substantial gain of 12.88%p when tested on OLID. This consistent improvement across different datasets highlights the enhanced generalization capability of models trained with our data selection strategy.

6.2 Quantitative Analysis

Model		SBIC	DYNA	ETHOS	OLID	ToxiGen
Llama3.1-8B		73.23	66.67	65.77	90.72	66.49
GPT-3.5-turbo		84.23	75.59	73.00	78.95	74.47
GPT-4o		81.32	83.19	74.00	81.55	70.73
GPT-4o-mini		76.40	**83.50**	72.00	56.47	67.53
S	BERT	87.56	73.47	76.59	90.35	64.08
	HateBERT	84.66	59.54	68.82	83.97	66.59
	Roberta	**84.95**	62.02	72.08	82.97	74.37
O	BERT	61.49	55.57	61.06	**97.28**	39.63
	HateBERT	58.78	51.75	61.26	95.57	62.10
	Roberta	59.25	55.92	60.84	95.63	63.17
E	BERT	66.59	61.31	71.99	72.39	48.22
	HateBERT	62.35	60.52	78.00	71.07	66.83
	Roberta	61.34	60.13	**78.37**	70.00	**66.86**

Table 4: F1 comparison between LLMs and traditional language models. S, O and E refer to models trained on the SBIC, OLID and ETHOS datasets, respectively. The highest score is highlighted in **bold**.

In Table 4, we conduct additional experiments using RoBERTa and HateBERT to assess the robustness of our CONELA strategy. We compare its performance against large language models (LLMs) using zero-shot inference and observe that our strategy often outperforms them.

For instance, traditional models trained on the SBIC dataset, such as BERT, exhibit superior performance with an F1 score of 87.56, outperforming LLMs like GPT-3.5-turbo (84.23) and GPT-4o (81.32) on the same dataset. This trend is further observed for models like RoBERTa, which achieves an F1 score of 84.95 on SBIC, surpassing LLMs like GPT-4o-mini (76.40). For the DynaHate dataset, GPT-4o achieves an F1 score of 83.19, which notably surpasses traditional models such as BERT (73.47) and RoBERTa (62.02). We notice that the GPT-4o's performance is comparable to that of traditional models on several datasets. For example, GPT-4o achieves 83.19 on DynaHate and 81.55 on OLID, which are competitive with or surpass the performances of BERT and RoBERTa in some cases. These results indicate that our CONELA strategy, particularly when applied to in-domain datasets like SBIC, improves performance and generalization compared to LLMs, confirming the effectiveness of our approach.

Model	SBIC	DYNA	ETHOS	OLID	ToxiGen
base	58.93	58.47	68.56	59.51	40.06
CONELA	**66.59**	61.31	71.99	**72.39**	48.22
CONELA-s	61.84	61.54	73.18	69.08	47.04
CONELA-s-ens	61.55	**62.12**	**77.00**	49.53	**66.59**

Table 5: F1 results of CONELA-s trained with ETHOS. The highest score is highlighted in **bold**.

6.3 The Data Scarcity Case

Our analysis reveals that the ETHOS training dataset shows less performance improvements

due to its limited data resources compared to other learning datasets. Addressing this limitation, we implement the disagreement loss in Equation (4), which results in notable performance enhancements reported in Table 5. Specifically, the CONELA-s-ens model achieves the highest F1 score of 77.00 on the ETHOS dataset, significantly surpassing both the base model (68.56) and other CONELA variants. This result highlights the effectiveness of the ensemble approach in improving model performance when trained on small datasets like ETHOS, where data scarcity is a challenge.

Rather than eliminating the ETL-non-consensual and HTL-non-consensual components, we use an ensemble technique to dynamically assign weights, leveraging useful information while mitigating negative impacts. This is particularly evident in the ToxiGen dataset, where the CONELA-s-ens model achieves 66.59. The combination of Equation (3) and the ensemble method yields substantial performance improvements, addressing data scarcity in the ETHOS dataset and enhancing model performance across diverse hate speech categories. This approach proves effective in scenarios with limited training data, offering a more adaptive solution for hate speech detection tasks.

6.4 Linguistic Analysis

We perform a linguistic analysis to investigate the differences between the instances retained by CONELA and those discarded according to our strategy. Using the NLTK library, we extract features such as token count, unique token ratio, noun ratio, verb ratio, adjective ratio, sentence length, average word length, and unique lemma ratio. We apply the Mann-Whitney U test (Mann and Whitney, 1947; Wilcoxon, 1992) to compare these features between the two datasets.

The analysis shows clear differences in token count and unique token ratio that highlight variations in lexical diversity between the two datasets. The verb and adjective ratios differ as well and suggest that discarded instances tend to use these parts of speech more often. In contrast, features like noun ratio and average word length show no significant differences. Sentence length and unique lemma ratio vary noticeably and reflect differences in sentence structure and lexical variety.

These findings confirm that the two datasets exhibit clear linguistic differences, particularly in lexical diversity and sentence complexity. This suggests that the discarded instances may introduce noise into the model, making them more challenging for the model to learn from effectively.

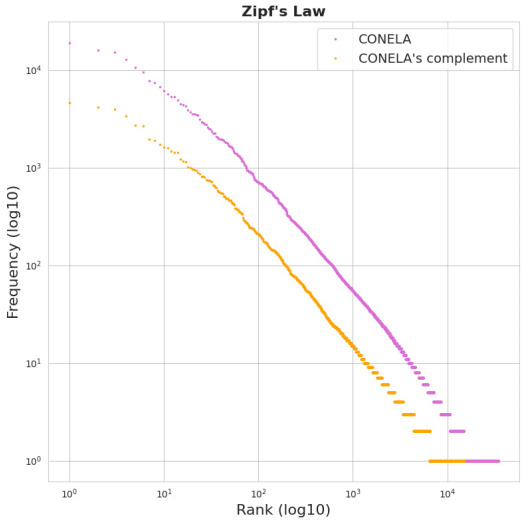


Figure 3: Word frequency distribution following Zipf's law for the CONELA dataset and its complement. The plot demonstrates the relationship between word rank and frequency on a log-log scale

6.4.1 Linguistic Diversity

We visualize the word frequency distribution in Figure 3 based on Zipf's law (Zipf, 2016)—the word frequency in a natural language corpus is inversely proportional to its rank. The ratio of the top 100 most frequent words in the CONELA dataset is 49.92% whereas it is 50.75% in the complement dataset. Additionally, the Type-Token Ratio (TTR) for CONELA is 6.16% whereas it is 10.12% for the complement dataset.

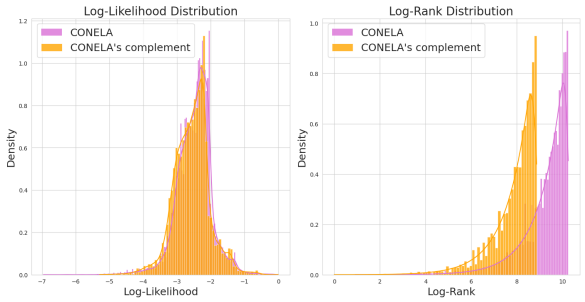
While the difference in the ratio of top 100 words between the two datasets is minimal, the higher TTR in CONELA's complement indicates greater lexical diversity in the discarded instances. This suggests that instances with more varied vocabulary tend to be more complex and nuanced; in other words, they are harder for both humans and models to classify as offensive. The linguistic diversity in CONELA's complement likely contributes to the model's difficulty in learning from these instances. This finding underscores the challenge of training models on lexically diverse and ambiguous data and highlights the need for more advanced strategies to handle such instances in natural language processing tasks.

6.4.2 Kolmogorov-Smirnov Test Results

The Kolmogorov-Smirnov test is a non-parametric statistical method used to assess whether two

Condition	SBIC (ID)	DYNA (OOD)	ETHOS (OOD)	OLID (OOD)	ToxiGen (OOD)	Average
*Baseline (100% train)	85.52	72.21	72.47	83.64	61.67	75.10
w/o EtL Non-Consensual & HtL Non-Consensual (CONELA)	**87.56**	**73.47**	76.59	**90.35**	**64.08**	**78.41**
w/o All Non-Consensual	86.56	72.47	75.59	89.35	63.08	77.41
w/o EtL Non-Consensual	85.53	72.24	71.87	84.23	61.77	75.13
w/o HtL Non-Consensual	86.63	72.57	**76.62**	89.29	63.06	77.63
w/o AtL Non-Consensual	86.06	72.34	72.75	85.83	63.06	76.01
w/o HtL Non-Consensual & AtL Non-Consensual	85.69	72.10	72.08	84.50	62.78	75.43
w/o EtL Non-Consensual & AtL Non-Consensual	86.41	72.18	75.22	88.42	62.60	76.97

Table 6: Performance Comparison of BERT uncased Model Trained on the SBIC Dataset Across 7 Categorized Datasets and Conditions. The F1 scores are compared to a baseline; scores surpassing the baseline are highlighted in bold. Standard deviations are provided next to each score. The ID column represents the dataset used for training.



(a) Log-Likelihood (b) Log-Rank

Figure 4: Comparison of Log-Likelihood and Log-Rank distributions between CONELA and CONELA's complement datasets.

datasets follow the same distribution. As shown in Figure 4, both the Log-Likelihood and Log-Rank distributions reveal significant differences between the CONELA and CONELA's complement datasets. The particularly large value for the Log-Rank statistic suggests the presence of distinct patterns in the data, which likely influence the model's learning process. These differences in distribution highlight the challenges the model faces when learning from the complement data, where lexical and structural diversity appear to be more pronounced.

7 Ablation Studies

We conduct ablation studies to analyze the impact of different data categories on model performance. Table 6 presents the results of systematically excluding subsets of non-consensual data derived from model training dynamics (EtL, AtL, HtL).

The results show that excluding both EtL Non-Consensual and HtL Non-Consensual data (CONELA condition) leads to the best overall performance. In this case, the model achieves an average F1 score of 78.41, showing a clear improvement over the baseline of 75.10. This condition also generalizes well to OOD datasets with particularly strong results on OLID.

When excluding HtL Non-Consensual data alone, the model sees notable gains especially on ETHOS where the F1 score reaches 76.62. This suggests that HtL contributes meaningfully to improved robustness. On the other hand, excluding EtL Non-Consensual alone provides only modest benefits with results showing smaller and less consistent improvements. Removing AtL Non-Consensual data leads to slight performance increases but falls short of the gains observed in the CONELA condition. For example, while the average score improves compared to the baseline, it remains below what is achieved by excluding both EtL and HtL.

These findings highlight that HtL Non-Consensual data plays a more critical role in enhancing performance, while EtL and AtL have limited impact. The combination of EtL and HtL exclusions stands out as the most effective strategy for training models that perform well both ID and OOD datasets.

8 Conclusion

We notice the inherent noise present in hate speech datasets largely due to the subjective nature of annotations. We tackle this issue by enhancing the quality of the dataset. Based on the initial empirical analysis, we have identified major factors that contribute to the degradation of model performance by developing a datamap that illustrates the agreement level among annotators across three categories: easy-to-learn, hard-to-learn, ambiguous. Our findings have suggested that sentences categorized as easy-to-learn, while having low-agreement among human annotators–indicating instances where human judgment finds difficulty, yet model does not–constitute poor-quality data. By excluding posts that annotators have different perspectives, we have observed a notable improvement of model performance. Our experimental results

have confirmed that the proposed data selection approach CONELA improves model performance substantially.

Limitations

We propose a data refinement strategy that concurrently considers model confidence and the degree of human annotation agreement, promoting learning through meticulous analysis. However, the reliability of human annotations remains a critical concern. The degree of human agreement is calculated as the mean of the values provided by multiple annotators. As a result, a single extreme outlier can significantly skew the average even when there is uniformity among other annotators. This sensitivity to outliers highlights the need for robust methods to manage such cases. Nonetheless, our data selection strategy demonstrates that excluding EtL-Non-Consensual and HtL-Non-Consensual instances facilitates improvements in model performance particularly in out-of-distribution generalization.

Ethical Considerations

Hate speech detection systems face challenges in balancing freedom of expression and effective moderation as overly aggressive detection risks censorship while misclassifications can lead to unjust outcomes. These systems may perpetuate societal biases, struggle with context, nuance, and cultural sensitivity, and raise concerns about privacy, transparency, and power dynamics. Moreover, evolving language and the need for scalability versus accuracy complicate their development, requiring continuous updates to address these issues responsibly. Addressing these ethical considerations requires ongoing collaboration between technologists, ethicists, policymakers, and diverse community representatives to ensure hate speech detection systems are effective, fair, and respectful of fundamental rights.

Acknowledgments

This research was supported by the NRF grant (RS2023-00208094) and the AI Graduate School Program at Yonsei University (RS-2020-II201361) funded by the Korean government (MSIT).

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. Sharedcon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10444–10455.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Venkateswarlu Bonta, Nandhini Kumaresh, and Naulegari Janardhan. 2019. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *WebSci*, pages 11–20. ACM.

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2020. Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472.

Junfan Chen, Richong Zhang, Junchi Chen, and Chunming Hu. 2024. Open-set semi-supervised text classification via adversarial disagreement maximization. In *ACL (1)*, pages 2170–2180. 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 *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *WSDM*, pages 231–240. ACM.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *EMNLP (1)*, pages 345–363. Association for Computational Linguistics.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *RANLP*, pages 260–266. INCOMA Ltd.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *ACL (1)*, pages 3309–3326. Association for Computational Linguistics.

Shi Yin Hong and Susan Gauch. 2023. Improving cross-domain hate speech generalizability with emotion knowledge. In *PACLIC*, pages 282–292. Association for Computational Linguistics.

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *J. Big Data*, 6:27.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *COLING*, pages 6667–6679. International Committee on Computational Linguistics.

Jan Kocon, Joanna Baran, Kamil Kanclerz, Michal Kajstura, and Przemyslaw Kazienko. 2023. Differential dataset cartography: Explainable artificial intelligence in comparative personalized sentiment analysis. In *ICCS (1)*, volume 14073 of *Lecture Notes in Computer Science*, pages 148–162. Springer.

Yongchan Kwon and James Zou. 2023. Data-oob: Out-of-bag estimate as a simple and efficient data value. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 18135–18152. PMLR.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, pages 6402–6413.

Seanie Lee, Dongho Kim, Hwijeen Kim, Jamin Shin, and Sung Ju Hwang. 2023. Adm: Adversarial disagreement maximization for robust text classification. *arXiv preprint arXiv:2305.14913*.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). In *SemEval@ACL*, pages 2304–2318. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *EMNLP (1)*, pages 10528–10539. Association for Computational Linguistics.

Changchun Li, Ximing Li, and Jihong Ouyang. 2021. Semi-supervised text classification with balanced deep representation distributions. In *ACL/IJCNLP (1)*, pages 5044–5053. 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. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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:e0221152.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI*, pages 14867–14875. AAAI Press.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of nli models. *arXiv preprint arXiv:2106.03020*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, pages 11048–11064. Association for Computational Linguistics.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 21–30.

Debaditya Pal, Kaustubh Chaudhari, and Harsh Sharma. 2022. Combating high variance in data-scarce implicit hate speech classification. In *TENCON 2022-2022 IEEE Region 10 Conference (TENCON)*, pages 1–4. IEEE.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *NeurIPS*, pages 20596–20607.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Trans. Assoc. Comput. Linguistics*, 7:677–694.

Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *NeurIPS*.

Md Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mucahid Kutlu, and Matthew Lease. 2021. An information retrieval approach to building datasets for hate speech detection. *arXiv preprint arXiv:2106.09775*.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *NAACL-HLT*, pages 175–190. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*, pages 5477–5490. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *EMNLP*, pages 973–981. ACL.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2023. Learning from noisy labels with deep neural networks: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 34(11):8135–8153.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *EMNLP (1)*, pages 9275–9293. Association for Computational Linguistics.

Ilan Tal and Dan Vilenchik. 2022. HARALD: augmenting hate speech data sets with real data. In *EMNLP (Findings)*, pages 2241–2248. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *ICLR (Poster)*. OpenReview.net.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *J. Artif. Intell. Res.*, 72:1385–1470.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL/IJCNLP (1)*, pages 1667–1682. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language - what does it actually look like and why are we not getting there? In *NAACL-HLT*, pages 576–587. Association for Computational Linguistics.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

Hai-Ming Xu, Lingqiao Liu, and Ehsan Abbasnejad. 2022. Progressive class semantic matching for semi-supervised text classification. In *NAACL-HLT*, pages 3003–3013. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *NAACL-HLT (1)*, pages 1415–1420. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *ICLR (Poster)*. OpenReview.net.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *EMNLP (Findings)*, pages 395–406. Association for Computational Linguistics.

Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *ACL (1)*, pages 6120–6130. Association for Computational Linguistics.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. In *ACL/IJCNLP (1)*, pages 7158–7166. Association for Computational Linguistics.

Zhi-Hua Zhou and Ming Li. 2010. Semi-supervised learning by disagreement. *Knowl. Inf. Syst.*, 24(3):415–439.

George Kingsley Zipf. 2016. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Ravenio books.

A Dataset Details

- **SBIC** (Sap et al., 2020) dataset provides a rich collection of social media posts annotated with structured implications about a wide range of demographic groups.
- **OLID** (Zampieri et al., 2019) is a hierarchical dataset that aims to classify offensive texts on social media into various categories and targets, which is collected on Twitter.
- **ETHOS** (Mollas et al., 2022), derived from YouTube and Reddit comments, offers both binary and multi-label classification challenges, showcasing the varied dimensions of hate speech across different platforms.
- **DYNAHATE** (Vidgen et al., 2021) employs a human-and-model-in-the-loop process for dynamically generating datasets over four rounds of dynamic data creation.
- **ToxiGen** (Hartvigsen et al., 2022) presents a large-scale machine-generated dataset focused on adversarial and implicit hate speech detection, leveraging advanced language models for data generation.

B Word Analysis of CONELA and CONELA's Complement

In our analysis, we explore the linguistic patterns inherent, subdividing it into 'CONELA' and 'CONELA's complement' based on the degree of human agreements, with t-SNE (Van der Maaten and Hinton, 2008). We extract the top keywords from each subset to understand the thematic content and linguistic intensity of the discussions.

The CONELA dataset is marked by terms that often relate to societal and racial topics, as well as common discourse elements. Examples of the top keywords in this dataset include:

- **Race-related terms**: 'black', 'white', 'jews', 'person'
- **General terms**: 'people', 'know', 'does', 'did', 'common'
- **Contextual terms**: Frequently neutral in isolation, but gaining significance within broader discussions, such as 'say'.

This dataset reflects discussions that often involve sensitive social and racial issues, with more factual or neutral expressions, often emphasizing descriptive or explanatory tones.

The CONELA's complement dataset, on the other hand, is characterized by more explicit and emotionally charged language. The top keywords in this subset include:

- **Explicit content**: 'f*cking', 'f*ck', 'n*gga', 'hate'
- **General terms**: 'don', 'just', 'people', 'say'
- **Emotive expressions**: 'like', 'want', 'hate'

This group is more representative of overt and emotionally intense discourse. The use of explicit language and emotionally charged terms reflects a higher degree of hostility and aggression, which can make the content more sensitive and challenging to moderate.

The CONELA dataset tends to focus more on factual and contextually significant terms, often revolving around social issues, with less emotionally intense language. In contrast, the CONELA's complement dataset features more explicit and emotionally driven terms, reflecting a greater degree of hostility and aggression. This contrast highlights the linguistic differences between the two datasets, with CONELA demonstrating a more neutral and descriptive tone, while CONELA's complement is characterized by heightened emotional intensity and explicit language. These differences underscore the challenges in identifying and moderating harmful content, especially when context plays a crucial role in determining its severity.

C Experimental Setup

Our experimental framework leverages the BERT-based architectures such as BERT-uncased and HateBERT, along with domain-specific models like RoBERTa, to address the task of implicit hate speech detection across various datasets including SBIC, OLID, DYNAHATE, ETHOS, and ToxiGen. The training configurations are meticulously set to ensure consistency and reproducibility across evaluations.

- **Hardware Configuration:** All models are trained on systems equipped with NVIDIA RTX4090 GPUs with operations performed on CUDA-enabled devices unless specified otherwise.
- **Training Parameters:** The models are trained for up to 10 epochs, with a learning rate of 5×10^{-6} and a batch size of 30. These parameters were selected to balance training speed and system capabilities.

This setup enables rigorous analysis of model performance across varied and complex hate speech scenarios, ensuring that findings are robust and broadly applicable.

C.1 Datamap Setup

The configuration for the data mapping via training dynamics is outlined as follows. The settings were chosen to optimize the performance of the BERT model in classifying textual data into predefined categories based on their ease of learning:

- Learning Rate (LR): 5×10^{-6}
- Number of Training Epochs: 6
- Patience for Early Stopping: 3
- Model Name: `bert-base-uncased`
- Random Seed: A random seed was used to ensure reproducibility of the results.

These parameters were set to fine-tune the model on the dataset, considering both the complexity of the language understanding task and the computational efficiency.

D Prompt-Based Binary Classification

English	Prompt Template
Hate Speech Detection:	You are a hate speech detection model. Analyze the given text. Respond with "offensive" if the text contains hate speech or offensive content, and "not offensive" otherwise. Respond in English only. Input text: [TEXT]

Table 7: Hate Speech Detection Prompt Template

In this experiment, binary classification was conducted using the models LLaMA 3.1, GPT-4o, GPT-3.5-turbo, and GPT-4o-mini, based on the provided prompt (see Table 7). For the test set, a uniform random sample of 100 instances was drawn from the original dataset for evaluation purposes.