EASY: Enhanced Analysis Approach for Implicit Hate Speech Yield – Bridging Human Insight and Algorithmic Precision in Social Media Discourse

Anonymous ACL submission

Abstract

The spread of abusive speech on social media influenced by genders, religions and context is a persistent challenge for hate speech detection. Previous researches focused on model-centric approaches and often overlooked the differences in how models and humans interpret offensive data. We propose a different approach that takes into account this prediction discrepancy for detecting hate speech more accurately. We advocate for the exclusion of sentences from the training dataset that are easily classified as hate speech by models but are challenging for humans. Our experiments on various datasets confirms that it is better to consider human agreement levels during the data preprocessing to improve the model generalization. This deviation underlines the unique challenges of hate speech domains, emphasizing the importance of datasets that reflect both model interpretations and human consensus. The analysis highlights the significance of a balanced dataset preparation approach to enhance the effectiveness and reliability of hate speech detection.

1 Introduction

Warning: This paper discusses and contains content that can be offensive or upsetting.

Implicit hate speech detection depends significantly on subjectivity, where the perception of whether content is hateful can vary depending on the context. A seemingly innocuous statement, such as "Honestly, I hate college but one benefit is you get a good sleeping schedule; it's 3:20 now, and I'm not even feeling tired," illustrates the complexity of identifying content that may carry underlying hateful sentiments based on the perspective of the reader. Early research faced challenges in handling context, leading to studies that relied on lexicon-based approaches to identify patterns in words or phrases (Ding et al., 2008; Lee et al., 2018; Bonta et al., 2019). More recent efforts have focused on generalizing implicit hate speech detection (Ludwig et al., 2022; Kim et al., 2022) and have introduced methods aimed at enhancing out-of-domain performance. Furthermore, it has been emphasized that hate datasets should not be judged based on a single opinion but rather annotated to reflect diverse viewpoints (Assimakopoulos et al., 2020). In the realm of offensive language, it is suggested that removing data deemed incorrect may overlook the critical roles of subjectivity, bias, and ambiguity (Leonardelli et al., 2021). It is also argued that even incorrectly annotated data should be preserved for its potential learning value (Leonardelli et al., 2023). Research is also moving towards approaching hate data from a human-centric perspective Kocoń et al. (2021), contrasting with attempts to generalize from a data-centric viewpoint Ramponi and Tonelli (2022). While recent research has extensively explored data quality, studies focusing on the interplay between human subjectivity and model understanding remain scarce. Our goal is to incorporate both model-centric and human-centric perspectives to measure generalization performance within the hate domain, conduct in-depth analysis from the viewpoint of subjectivity—an essential characteristic of offensive language detection—and propose related methodologies. Following the approach outlined in Swayamdipta et al. (2020), we plan to observe training dynamics and experimentally divide a dataset into three groups (easy, ambiguous, hard) based on confidence and variability, aiming to enhance the detection process.

The contributions of this study are threefold:

1. It underscores the critical role of data preprocessing within hate speech domains, advocating for methodologies that enhance model learning outcomes through refined dataset preparation.

2. It validates the proposed approach through

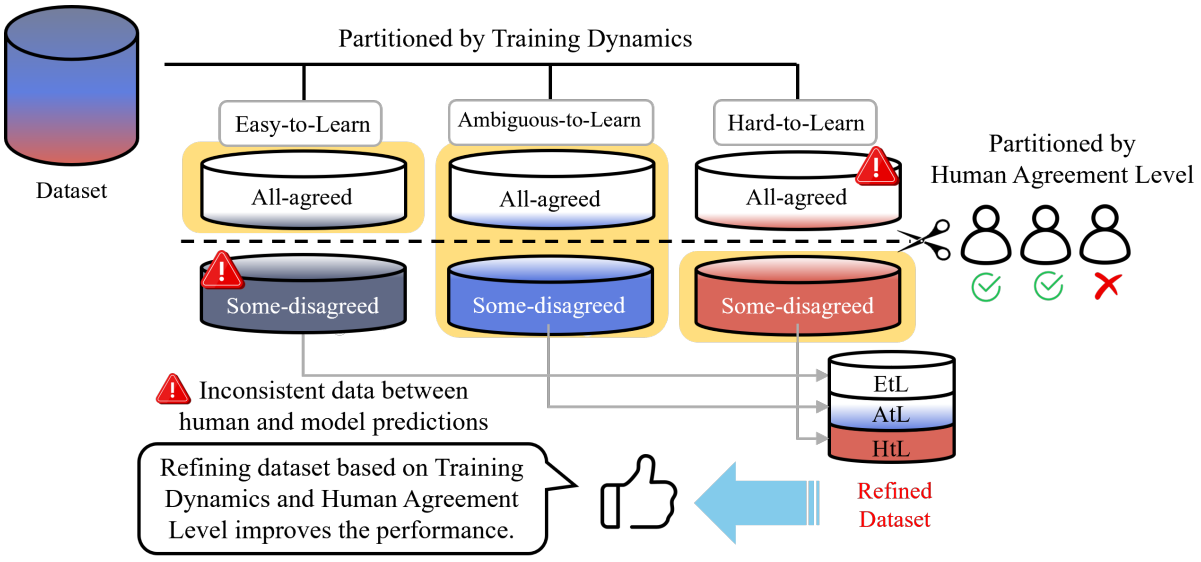

Figure 1: Overview of the methodology for hate speech dataset analysis: The approach categorizes data into easy, hard, and ambiguous sets based on model perspectives, incorporating human agreement levels to refine these into six distinct categories. The method emphasizes discarding categories without model-human consensus to enhance data quality.

rigorous cross-dataset evaluations, illustrating the robustness of our methodology and its potential applicability in diverse contexts.

3. It highlights the necessity of incorporating a human-centric perspective in the analysis of datasets , particularly those that models find easy to interpret but humans find challenging, thus ensuring a more effective and empathetic framework for hate speech detection.

2 Related Works

2.1 Training Dynamics

Recent works have delved into the challenge of detecting online toxicity, acknowledging that while subjectivity in data labels introduces complexity, efforts have been made to mitigate bias (Garg et al., 2023). These studies provide evidence that models trained on data characterized as easy data can often outperform those trained solely on more complex datasets (Hase et al., 2024). This revelation underscores the premise of our research and suggests that models which focus on data that is more readily learnable could exhibit enhanced robustness. Such a perspective endorses a refined strategy for data selection that considers not just the challenge posed by a dataset but also the potential for learning efficacy from simpler data configurations. Swayamdipta et al. (2020) adopts training dynamics to chart the learning trajectories of models across various datasets. Central to this approach are the metrics of *confidence* and *variability*. Confidence C quantifies the model's certainty in its prediction for a specific data point, while variability V monitors the fluctuation in this confidence over training epochs. These concepts are formally defined as:

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

$$V = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(P(y_i|x_i) - C)^2}.$$

Here, $P(y_i|x_i)$ is the probability assigned by the model to the correct label y_i for data point x_i over N epochs. These metrics assist in differentiating data points that are *Easy-to-Learn(EtL)*, *Hard-to-Learn(HtL)*, and *Ambiguous-to-Learn(AtL)*, facilitating a focused method for improving dataset integrity and, consequently, model performance. Additionally, Figure 2 presents a data map generated using the SBIC dataset to visualize the training dynamics, illustrating how data points are distributed among these categories.

2.2 Refining Natural Language Models through Annotator Insights

The quality of annotated data in supervised learning, especially for tasks like offensive language

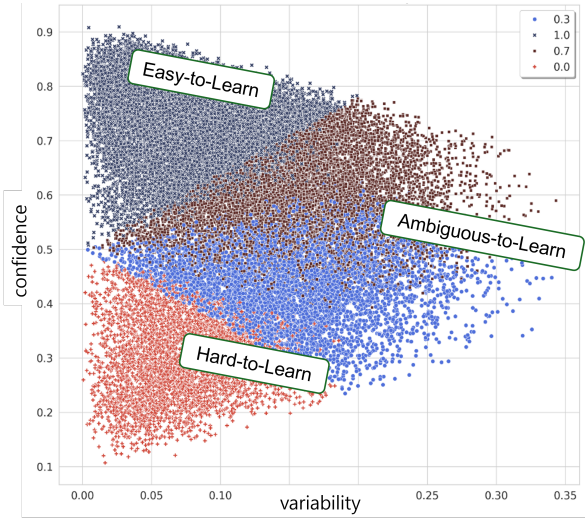

Figure 2: Data Map of the SBIC dataset for hate speech detection. This map represents training dynamics, where the x-axis denotes variability and the y-axis denotes model confidence. The color coding of points reflects the prediction accuracy such that low variability and high confidence suggest Easy-to-Learn regions, and the opposite implies Hard-to-Learn areas. The high variability regions are categorized as Ambiguous-to-Learn.

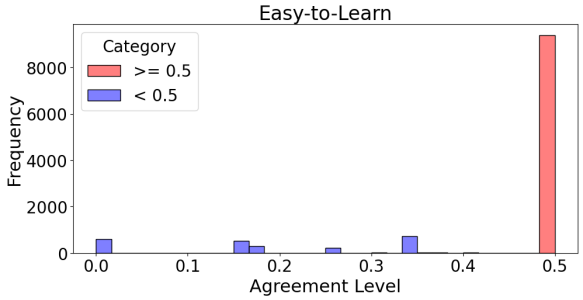

Figure 3: Bar plots illustrating the distribution of agreement levels across the categories of easy-to-learn within the SBIC dataset.

detection, is critically influenced by the nuances of human annotation. Studies reveal that both the inherent bias of crowd workers and the variability in annotator agreement significantly impact model performance and generalization (Leonardelli et al., 2021). Recently, Leonardelli et al. (Leonardelli et al., 2023) highlight the importance of integrating annotator disagreement to capture the diverse interpretations of offensive language, enriching model training with a broader spectrum of human judgment (Leonardelli et al., 2021). This perspective is echoed by Leonardelli et al. (2023), who argues that removing or altering annotations within the data is detrimental, potentially stripping the dataset of its rich diversity of opinion and expression. Leonardelli et al. (2021) explore how biases in crowdsourced annotations, driven by personal opinions, can skew data quality, advocating for mitigation strategies that enhance dataset integrity and, consequently, model reliability. Together, these insights underscore the dual necessity of accounting for linguistic discrepancies and mitigating annotation biases, suggesting a paradigm where nuanced annotation analysis becomes central to developing robust natural language models.

3 Method

Our proposed methodology, termed EASY (Enhanced Analysis Approach for Implicit Hate Speech Yield), integrates both model perspectives and human subjectivity in analyzing hate speech data. We employ a training dynamics approach on datasets that include human agreement levels, categorizing data into three primary types: *EtL, HtL*, and *AtL* from a model's perspective. To further refine this classification, we incorporate human judgment by subdividing these categories into 'Consensual' and 'Non-Consensual' regions based on the agreement level calculated using the formula $|0.5 - \text{offensiveYN}|$. This measure assesses the perceived offensiveness by human annotators, leveraging indicators like 'offensiveYN'.

We first define Easy-to-Learn (EtL) as follows, with the definitions for the remaining categories provided in Table 2.

EtL Consensual: Sections where both models and humans agree on ease of classification. To understand the impact of data imbalance(Padurariu and Breaban, 2019), we explored training on 25%, 50%, and 75% of this data. As shown in Figure 3, the distribution of data across categories indicates a significant imbalance, with 9461 instances in **EtL Consensual**. We conducted experiments with these varying proportions to examine the effects of data imbalance on model performance.

EtL Non-Consensual: This category includes sentences that are considered easy by the model but difficult for humans to classify. Specifically, these are cases where the sentences may involve implicit hate speech, making their classification ambiguous even for humans. For instance, the sentence "How do you stop a baby from crawling in circles? You nail its other hand to the floor." could be perceived as dark humor, which some might

Dataset	Train	Refined Train (*ours*)	Test
SBIC	35,424	33,077	4,691
OLID	19,826	19,352	2,479
ETHOS	798	699	100
DynaHate	-	-	4,120
ToxiGen	-	-	8,960

Table 1: Datasets used in the experiments. Training involved only SBIC, OLID, and ETHOS, which include human agreement levels. Note that ToxiGen and DynaHATE were not used as training datasets because they do not include human agreement levels.

find amusing while others might see it as offensive. Such implicit hate speech requires a nuanced understanding of context, suggesting that these sentences, confidently classified as easy-to-learn by the model, might actually reflect biases that warrant further investigation. Thus, we removed these sentences to explore how their exclusion affects model performance.

Our experimental setup involves fine-tuning serveral pre-trained models, and we analyze the results across various categories using multiple models to ensure robust findings. The methodology and its comprehensive framework are depicted in Figure 1, illustrating the intersection of machine learning precision and human interpretative complexity. This approach aims to enhance the model's accuracy by aligning it more closely with nuanced human insights into what constitutes hate speech.

4 Experimental Results

4.1 Datasets

We conduct binary text classification to delve into the nuances of implicit hate speech datasets. For training, we utilize datasets that incorporate human agreement levels, specifically the Social Bias Inference Corpus (SBIC), Offensive Language Identification Dataset (OLID), and Online Hate Speech Detection Dataset (ETHOS). These datasets are annotated with agreement scores ranging from 0.0 to 1.0, reflecting a spectrum of human consensus on the offensiveness of content. For testing, we employ the DYNA HATE and ToxiGen datasets. Detailed information about the sizes of these datasets can be found in Table 1.

- **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, making it a valuable resource for understanding the multifaceted nature of offensive language.
- **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., 2020) introduces a novel approach to dataset creation by including examples specifically designed to challenge hate speech detection models, thus enhancing their adaptability and robustness.
- **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.

In this analysis, we refine our approach by preprocessing the datasets to focus solely on the posts (sentences) and their associated offensiveYN labels, which allows us to perform binary classification on whether content is considered hate speech or not. This preprocessing step ensures that we leverage only the most pertinent columns for our analysis, thus enhancing the relevance of our training and evaluation phases.

4.2 Baseline Experimental Setup

In our experimental framework, we employ several baseline models to establish a comprehensive understanding of performance across different architectures and setups. The primary models used are BERT [1] (Devlin et al., 2018; Saleh et al., 2023) and its specialized derivative, HateBERT [2] (Caselli et al., 2020), known for their effectiveness in processing natural language and detecting hate speech nuances. Additionally, we include domain-specific models like ToxiGen-RoBERTa to diversify our experimental insights. We conduct experiments with multiple seeds, ranging from three to five, to ensure the robustness and reproducibility of our results. The learning rate is set to 5e-6, with batch

[1] https://huggingface.co/google-bert/bert-base-uncased
[2] https://huggingface.co/GroNLP/hateBERT

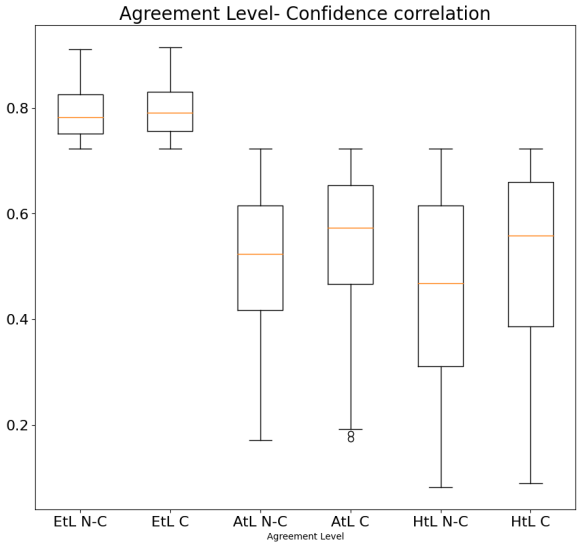

Figure 4: Correlation between agreement levels and model confidence and variability across EtL, HtL, and AtL categories within the SBIC dataset. The plot shows the agreement and confidence correlation. On the x-axis, 'N-C' represents Non-Consensual, while 'C' stands for Consensual.

sizes varying between 16 and 30 to optimize computational efficiency and learning dynamics. We utilize NVIDIA RTX4090 GPUs with a batch size of 30 per device. The labels in datasets featuring human agreement levels are processed as floating-point numbers ranging from 0.0 to 1.0. We utilize a thresholding approach where scores above 0.5 are classified as hate (1) and those below as not hate (0). For our evaluation metrics, we rely on Accuracy and the F1 score to assess both the precision and recall capabilities of our models comprehensively. Optimization is performed using the AdamW optimizer, which is noted for its effectiveness in fine-tuning large pre-trained models. Detailed configurations of our experimental setup, including hyperparameters and additional methodological nuances, are meticulously documented in the appendix A of this paper. This extensive setup allows us to conduct a deep analysis of how different configurations impact performance, particularly focusing on the interplay between model outputs and the subjective interpretation of hate speech by humans.

5 EASY: Enhanced Analysis Approach for Implicit Hate Speech Yield

5.1 EtL Analysis

One of the significant challenges within the hate speech domain is the generalization of implicit hate speech detection. Research is conducted using methods such as debiasing and contrast learning to address these issues (Badjatiya et al., 2019). We hypothesize that removing data considered easy by the model but challenging from a human perspective can improve performance levels without the need for direct fine-tuning of the model's architecture. Currently, we examine the actual differences between the Consensual and Non-Consensual regions of the EtL data trained on the SBIC using training dynamics. Both data groups are found to have confidence levels around 0.8 (figure 4).

5.1.1 Linguistic Patterns in EtL Dataset

In our analysis, we explore the linguistic patterns inherent in the Easy-to-Learn (EtL) dataset, subdividing it into 'EtL Consensual' and 'EtL Non-Consensual' based on the consensus level of annotators. We extract the top keywords from each subset to understand the thematic content and linguistic intensity of the discussions.

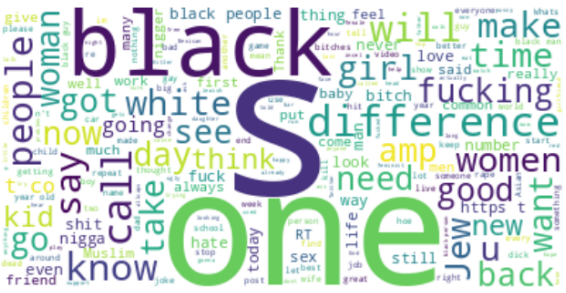

Figure 5: Words clouds for EtL Consensual Dataset

EtL Consensual The 'EtL Consensual' subset, which corresponds to content typically agreed upon as non-offensive or mildly offensive by both models and humans, exhibited frequent occurrences of everyday terms intertwined with explicit language. Figure 5 illustrates the top keywords, which include:

- **General terms**: 'one', 'people', 'day', 'time', 'see', 'go'
- **Explicit content**: 'f*cking', 'b*tch', 'f*ck'
- **Race-related terms**: 'black', 'white', 'Jew', 'black people'

- **Emotive expressions**: 'like', 'want', 'need', 'good'

This subset is characterized by a more direct and overt expression, reflecting clear stances or opinions.

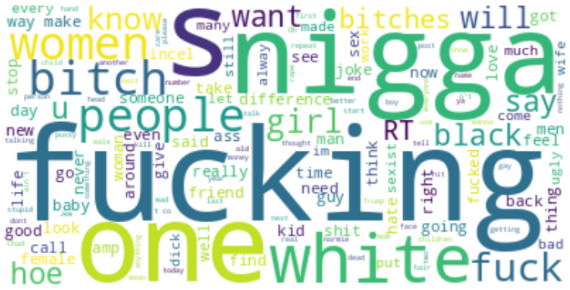

Figure 6: Words clouds for EtL Non-Consensual Dataset

EtL Non-Consensual Conversely, the 'EtL Non-Consensual' subset includes terms that often relate to sensitive societal topics, showcasing a broader spectrum of subjects and higher emotional intensity. The top keywords from Figure 6 include:

- **Explicit content**: 'f*cking', 'n*gga', 'b*tch', 'f*ck', 'shit', 'h*e', 'a*s'
- **General terms**: 'people', 'women', 'girl', 'man', 'kid'
- **Discriminatory language**: Often implicit through the context in which even commonplace words are used.
- **Calls to action or emotions**: 'want', 'need', 'hate', 'love'

This group highlights the complexities of defining hate speech, where the context or the presence of certain keywords escalates the sensitivity of the content.

5.1.2 Implications for Hate Speech Detection

This comparative study suggests the necessity for granulated categorization within hate speech datasets. By segmenting the datasets into more manageable sub-groups based on explicitness and societal sensitivity, we can fine-tune hate speech detection models for improved performance and better understanding of the complexities involved.

Moreover, our research advocates for incorporating a data selection process that factors in human subjectivity and annotator consensus. This strategy not only aids in reducing model bias but also enriches the models' capability to discern between overt and subtle forms of hate speech.

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

Table 2: Description of data categories used in the study.

5.2 In-depth Analysis

Our hypothesis posits that by considering both human-centric and model-centric perspectives during data preprocessing, the quality of the data improves, thereby enhancing the generalization performance for implicit hate speech detection. To this end, data initially classified from a model-centric standpoint is reprocessed to incorporate human-centric subjectivity. We conduct an extensive evaluation using various datasets and ablation studies to analyze the results, which are detailed further in the appendix A. The robustness of our hypothesis is tested by training on three datasets and testing on five different datasets. This approach not only reinforces the reliability of our findings but also allows us to test the applicability across different language models. We categorize the dataset that reflects subjectivity into six major groups. The characteristics and definitions of these categories are detailed in Table 2.

5.2.1 Performance Results

Overall, as presented in Table 3, EASY demonstrates performance improvements in most Out-of-Domain (OOD) tests. The BERT model, pretrained on the SBIC dataset, exhibits a maximum performance increase of 6.04%pt in the OLID dataset. Additional gains are observed in DynaHate by 3.99%pt and ToxiGen by 3.24%pt, confirming the impact on generalization performance. Testing on the smallest dataset we train, ETHOS, shows the most significant change, with a maximum improvement of 10.31%pt.

To validate our hypothesis further, we conduct tests using other language models, such as HateBERT and ToxiGen-RoBERTa. The results, generally showing improved or similar performance levels, are detailed in Table 4.

Furthermore, as indicated in Appendix A, adjusting the quantity of EtL Consensual data does

Test	Train					
	SBIC		OLID		ETHOS	
	baseline	*Ours*	baseline	*Ours*	baseline	*Ours*
SBIC	80.12 ± 1.4	79.77 ± 1.2	70.78 ± 4.7	**71.63** ± 3.0	53.29 ± 7.3	**65.81** ± 3.2
OLID	43.80 ± 9.5	**49.84** ± 2.9	90.03 ± 1.0	**92.83** ± 2.1	36.14 ± 9.0	**43.07** ± 9.0
ETHOS	63.05 ± 2.5	**63.16** ± 0.7	58.37 ± 4.7	**60.02** ± 0.8	68.81 ± 5.2	**71.87** ± 5.7
DynaHate	63.44 ± 5.7	**67.42** ± 2.3	66.85 ± 8.0	**68.47** ± 4.6	51.55 ± 2.9	**61.86** ± 1.5
ToxiGen	52.87 ± 8.8	**56.12** ± 6.3	30.77 ± 6.6	**61.65** ± 6.3	31.80 ± 12.1	**40.44** ± 9.0

Table 3: Performance Comparison: F1 Score Performance Comparison of BERT-uncased Trained on SBIC, OLID, ETHOS Dataset Across Different Datasets and Conditions

Model	Condition	SBIC (ID)	DynaHate (OOD)	ETHOS (OOD)	OLID (OOD)	ToxiGen (OOD)
HateBERT	*Baseline (100% train)	84.64 ± 0.3	60.79 ± 0.5	71.65 ± 1.1	68.38 ± 0.5	59.73 ± 0.3
HateBERT	**w/o EtL Non-Consensual**	**84.76** ± 0.2	**60.94** ± 0.4	**72.94** ± 0.8	67.52 ± 0.6	**60.41** ± 0.5
ToxiGen_Roberta	*Baseline (100% train)	84.98 ± 0.3	63.93 ± 1.4	73.12 ± 2.0	68.52 ± 1.2	75.15 ± 0.5
ToxiGen_Roberta	**w/o EtL Non-Consensual**	**85.19** ± 0.4	**64.14** ± 1.9	**74.48** ± 1.7	67.79 ± 1.0	**75.30** ± 1.0

Table 4: F1 Score Comparison Across Models and Conditions, Based on Training with the SBIC Dataset

not significantly affect performance, suggesting that removing EtL Non-Consensual data is more crucial. Despite comprising only 2,347 out of 35,424 data points in the SBIC dataset, the EtL Non-Consensual category shows the highest performance improvement. These results suggest that when models easily predict labels for data that humans find difficult to judge as offensive, it may indicate misdirection in model training. Therefore, enhancing the quality of the EtL-classified dataset can significantly impact the overall dataset quality, underscoring the importance of improving EtL data quality.

5.2.2 Ablation Study

In addition to testing our initial hypothesis, Appendix A.4, which presents the results of experiments that involved various combinations of the categories defined in Table 2. For example, the configuration without (w/o) EtL Non-Consensual, AtL Consensual, and HtL Consensual involved removing all datasets where human and model perspectives differ and training solely on the remaining data. The results predominantly showed a decline in performance. Specifically, removing only AtL Consensual, where both models and humans find the data difficult, and training without EtL Non-Consensual resulted in a maximum performance drop of up to 7.9 %pt. This suggests that data deemed difficult by both models and humans should be used as training material.

Moreover, training solely on EtL Consensual, where human and model perspectives align, also resulted in a slight decrease in performance. This indicates that datasets perceived as easy by models might mislead the training direction. Training only on AtL Consensual data led to a significant drop in performance, and excluding AtL Consensual from training also resulted in substantial performance decreases. This highlights the influence of these data when trained alongside other datasets. Through a series of 19 ablation studies, we have demonstrated that removing the EtL Non-Consensual category has the most significant impact on improving generalization performance.

5.3 Discussion

5.3.1 Implications of Dataset Quality

This research is predicated on the notion that performance in the task of implicit hate speech detection can be enhanced not through modifications to the model itself but by improving the quality of the dataset. Our research employs a data mapping methodology, classifying data into three areas based on human agreement levels: Easy-to-Learn (EtL), Hard-to-Learn (HtL), and Ambiguous-to-Learn (AtL). The analysis confirms that focusing on the quality of data within the EtL area is crucial.

However, we acknowledge ongoing concerns regarding the reliability of the human agreement level. Since the agreement level is an average of labels provided by multiple annotators, a single outlier can significantly alter the average, even if all other annotators agree. Additionally, poor performance in offensive data detection tasks may be attributed to a lack of contextual understanding

or inadequate grasp of English slang, which can similarly affect human annotators, leading to potential mislabeling if they do not fully comprehend a sentence. Therefore, attention must also be given to datasets where the agreement level is measured low (EtL Non-Consensual, AtL Consensual, HtL Consensual). As shown in Appendix A.4, training exclusively on datasets from each area reveals that training only with HtL Consensual results in the lowest performance, whereas removing HtL Consensual data (w/o HtL Consensual) does not significantly enhance performance. This suggests a need for further differentiation within datasets classified as AtL Consensual and HtL Consensual. For instance, analyzing the standard deviation of the agreement level might help determine whether the low level is due to a divergence of opinions among annotators or the extreme labeling by an individual, thus potentially extracting true HtL data. In conclusion, enhancing the dataset quality, especially by scrutinizing and refining data in low agreement areas, can significantly impact the performance and reliability of implicit hate speech detection models.

5.3.2 Insights from the HtL Category

As depicted in Figure 7, the data map of the HtL classified subset does not concentrate within a singular region. Instead, it exhibits a well-distributed spread across three distinct areas: high confidence with low variability, high variability, and low confidence with low variability. This distribution contrasts with the data map generated using the entire dataset (see Figure 2), where the delineation of regions is significantly more pronounced. The observed pattern aligns closely with the phenomena reported in existing literature (Swayamdipta et al., 2020). The presence of data points within the high confidence, low variability zone, yet categorized as HtL, intimates that these instances may not have exhibited strong confidence in the overall data map. Nevertheless, objectively, these could be considered easy-to-learn. This discrepancy suggests a need for further research to understand the underlying factors contributing to such classification anomalies.

5.3.3 Comparison with Other Domain

We extended our analysis to datasets outside the hate domain to ascertain the specificity of our approach's effectiveness. The SST dataset (Socher et al., 2013), when compared to the SBIC dataset, exhibited substantially less overlap in regions of

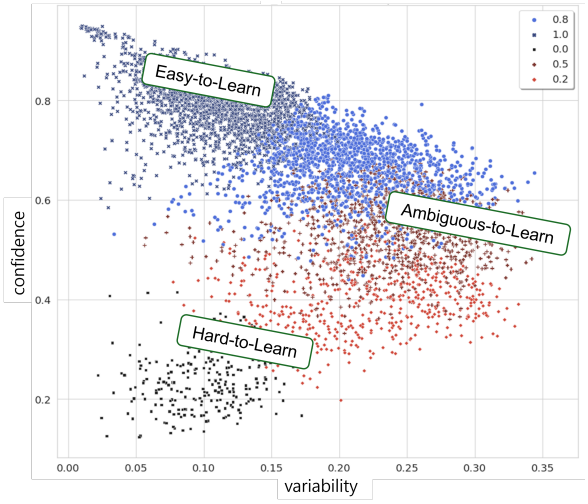

Figure 7: Hard-to-Learn Non-Consensual Datamap

ambiguity. This indicated a clearer demarcation between the different classifications in a general domain setting. We adopted the same methodology of segmenting the data into 'Consensual' and 'Non-Consensual' areas and evaluated the F1 scores accordingly. The removal of EtL Non-Consensual data in a non-hate domain, specifically when compared to the baseline, resulted in a performance decrement of 8.6%pt, confirming that our method's applicability is particularly pronounced within the hate domain.

6 Conclusion

In this study, we acknowledge the inherent noise present in hate speech datasets, largely attributable to the subjective nature of annotations. To address this issue, our approach has not been to refine the model but to enhance the quality of the dataset itself. Through empirical analysis, we identified factors contributing 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 suggest that sentences categorized as easy-to-learn, while having low-agreement among human annotators–indicating instances where human judgement finds difficulty, yet model does not–constitute poor-quality data. By training our classifier to disregard these sentences, we observed a notable improvement of model performance. Thus, we experimentally demonstrated that improving the quality of the dataset alone can improve model performance, which we expect will be useful for future research on dataset refinement.

Limitations

We propose a data refinement strategy that concurrently considers model confidence and human agreement, promoting learning through meticulous analysis. However, the reliability of the agreement level must also be addressed. The agreement level is the mean of the values labeled by multiple annotators, and thus, if even one annotator mistakenly provides an extreme outlier, the average can be significantly skewed, despite uniformity from other annotators. Nonetheless, our data cleaning strategy demonstrates that excluding 'EtL Non-Consensual' data can facilitate improvements in model performance. Investigating approaches to manage annotators with distinctly prominent opinions represents a promising avenue for future work.

References

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke Van Der Plas, and Albert Gatt. 2020. Annotating for hate speech: The maneco corpus and some input from critical discourse analysis. *arXiv preprint arXiv:2008.06222*.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The world wide web conference*, pages 49–59.

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.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

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. *arXiv preprint arXiv:2203.09509*.

Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. 2024. The unreasonable effectiveness of easy training data for hard tasks. *arXiv preprint arXiv:2401.06751*.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, and Yo-Sub Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113:22–31.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). *arXiv preprint arXiv:2304.14803*.

Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Cristian Padurariu and Mihaela Elena Breaban. 2019. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040. Association for Computational Linguistics.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

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. *arXiv preprint arXiv:2009.10795*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

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. *arXiv preprint arXiv:1902.09666*.

A Appendix

A.1 Details in Models

BERT: BERT, a transformer-based machine learning technique developed by Google, is considered to be well-suited for tasks in the hate speech domain. This suitability is largely due to the model being trained on extensive and diverse text corpora, including Wikipedia and BookCorpus, which provide a broad linguistic foundation for understanding complex language patterns and nuances.

HateBERT: HateBERT is a version of the BERT model, specifically trained to detect hate speech by leveraging over one million posts from banned communities on Reddit. Developed through a collaboration between the University of Groningen, the University of Turin, and the University of Passau, this model enhances the detection of offensive and harmful language across various platforms.

ToxiGen-RoBERTa: ToxiGen-RoBERTa[3] is a specialized adaptation of RoBERTa trained to identify toxic language. It has been fine- tuned to better understand the nuances and context of offensive and harmful language, making it highly effective for tasks involving hate speech detection and online safety monitoring.

[3]https://huggingface.co/tomh/toxigen_roberta

A.2 Data Preprocessing

In our study, we performed a standardized preprocessing procedure on multiple datasets including OLID, SBIC, DynaHate, ToxiGen, and ETHOS, to ensure the uniformity of data and to remove any elements that could potentially bias the outcomes of our hate speech detection models. The following steps were systematically applied to each dataset:

1. **Removal of Binary Indicators:** We removed the binary indicators (e.g., leading "b'") from strings which are typical artifacts from byte encoding.

2. **Elimination of User Mentions and URLs:** All user mentions (e.g., "@user") and URLs were stripped from the texts to prevent any personal identification and to focus solely on the content of the communications.

3. **Cleaning of Special Characters and HTML Tags:** Special characters, HTML tags, and emojis were removed to standardize the text for analysis. This includes stripping of transport and map symbols, flags, and other emoticons that do not contribute to the meaning of the text.

4. **Punctuation and Whitespace Normalization:** We retained only word characters, digits, single quotes, and whitespaces. All other punctuations were removed, and multiple spaces, tabs, and new lines were reduced to a single space to maintain text consistency.

Each dataset required minor adaptations in preprocessing to accommodate the specific format of the data source. For instance:

- In datasets with columns named ‘tweet‘ and ‘class‘, these were renamed to ‘post‘ and ‘offensiveYN‘ respectively, with the ‘offensiveYN‘ binary flag adjusted to 0 for non-offensive and 1 for offensive entries.

- In datasets like those with ‘comment‘ and ‘isHate‘, renaming and adjustments were similar, ensuring that labels are consistent across all datasets.

The processed data retained only the necessary columns, namely 'offensiveYN' and 'post', and each post was cleaned using the defined text cleaning function. This uniform preprocessing approach allows for a more accurate and fair comparison of model performance across different datasets. This comprehensive preprocessing ensures that the data used in our experiments are free from common textual noise and standard across all datasets, thus enhancing the reliability of our findings.

A.3 Experimental Setup

Our experimental framework leverages the BERT-based architectures such as BERT-uncased and HateBERT, along with domain-specific models like ToxiGen-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 8 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.
- **Evaluation Strategy:** Evaluation during training is conducted at the end of each epoch, and comprehensive validation is performed on multiple datasets to assess generalization across different contexts.
- **Optimization:** Gradient accumulation is utilized to stabilize training updates, with the AdamW optimizer managing weight updates. The training employs a linear warmup strategy over the initial steps to mitigate early large gradient updates.
- **Regularization:** Training includes techniques such as weight decay and learning rate decay to prevent overfitting on the training data.
- **Reproducibility:** To ensure the reproducibility of our results, we employ a total of five random seeds for initializing the training process. The reported results in all tables are the average outcomes across these seeds, providing a robust measure of model performance and stability.

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

A.3.1 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.

A.4 Dataset Specific Results

Condition	SBIC (OOD)	DynaHate (OOD)	Ethos (ID)	Olid (OOD)	Toxigen (OOD)
*Baseline (100% train)	0.611 ± 0.04	0.581 ± 0.03	0.879 ± 0.02	0.649 ± 0.01	0.520 ± 0.04
w/o EtL Non-Consensual	**0.616** ± 0.05	0.576 ± 0.04	0.805 ± 0.04	0.647 ± 0.01	**0.527** ± 0.06
w/o EtL Non-Consensual & EtL Consensual 75%	**0.692** ± 0.05	**0.670** ± 0.03	0.586 ± 0.07	0.389 ± 0.05	0.307 ± 0.11
w/o EtL Non-Consensual & EtL Consensual 50%	**0.676** ± 0.03	**0.649** ± 0.02	0.555 ± 0.04	0.363 ± 0.02	0.218 ± 0.07
w/o EtL Non-Consensual & EtL Consensual 25%	0.571 ± 0.18	0.507 ± 0.22	0.446 ± 0.21	0.308 ± 0.14	0.238 ± 0.19
EtL Consensual	0.012 ± 0.01	0.049 ± 0.04	0.055 ± 0.05	0.141 ± 0.13	0.208 ± 0.22
EtL Non-Consensual	0.011 ± 0.01	0.045 ± 0.05	0.041 ± 0.05	0.111 ± 0.14	0.173 ± 0.22
AtL Consensual	**0.727** ± 0.00	**0.690** ± 0.02	0.660 ± 0.01	0.478 ± 0.03	**0.569** ± 0.08
AtL Non-Consensual	0.068 ± 0.04	0.136 ± 0.06	0.335 ± 0.06	0.240 ± 0.05	0.276 ± 0.10
HtL Consensual	**0.732** ± 0.00	**0.709** ± 0.00	0.657 ± 0.00	0.508 ± 0.01	**0.655** ± 0.02
HtL Non-Consensual	**0.646** ± 0.03	**0.630** ± 0.03	0.496 ± 0.08	0.328 ± 0.06	0.216 ± 0.20
w/o EtL Consensual & HtL Non-Consensual & AtL Non-Consensual	**0.732** ± 0.00	**0.710** ± 0.00	0.658 ± 0.00	0.514 ± 0.00	**0.665** ± 0.00
w/o EtL Non-Consensual & HtL Consensual & AtL Consensual	0.007 ± 0.01	0.009 ± 0.02	0.008 ± 0.02	0.014 ± 0.03	0.006 ± 0.01
w/o HtL Non-Consensual	0.043 ± 0.02	0.111 ± 0.03	0.186 ± 0.04	0.259 ± 0.06	0.346 ± 0.09
w/o AtL Non-Consensual	0.034 ± 0.03	0.057 ± 0.05	0.116 ± 0.10	0.122 ± 0.10	0.132 ± 0.13
w/o HtL Non-Consensual & AtL Non-Consensual	0.022 ± 0.01	0.075 ± 0.03	0.101 ± 0.07	0.206 ± 0.10	0.141 ± 0.10
w/o EtL Non-Consensual & HtL Non-Consensual	0.160 ± 0.07	0.173 ± 0.07	0.253 ± 0.09	0.160 ± 0.08	0.120 ± 0.10
w/o EtL Non-Consensual & AtL Non-Consensual	0.308 ± 0.23	0.275 ± 0.22	0.287 ± 0.20	0.206 ± 0.14	0.195 ± 0.17
w/o EtL Non-Consensual & HtL Non-Consensual & AtL Non-Consensual	0.270 ± 0.14	0.277 ± 0.08	0.333 ± 0.13	0.290 ± 0.10	0.315 ± 0.20

Table 5: Performance Comparison of BERT uncased Model Trained on the ETHOS Dataset Across 19 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.

Condition	DynaHate (OOD)	ETHOS (OOD)	OLID (ID)	SBIC (OOD)	ToxiGen (OOD)
*Baseline (100% train)	0.668 ± 0.08	0.584 ± 0.05	0.920 ± 0.01	0.708 ± 0.05	0.308 ± 0.07
w/o EtL Non-Consensual	**0.685** ± 0.05	**0.600** ± 0.01	**0.928** ± 0.02	**0.716** ± 0.03	**0.617** ± 0.06
w/o EtL Non-Consensual & EtL Consensual 75%	0.635 ± 0.15	0.532 ± 0.21	0.447 ± 0.10	0.637 ± 0.19	**0.460** ± 0.21
w/o EtL Non-Consensual & EtL Consensual 50%	**0.696** ± 0.02	0.570 ± 0.12	0.464 ± 0.05	0.607 ± 0.21	**0.535** ± 0.05
w/o EtL Non-Consensual & EtL Consensual 25%	**0.703** ± 0.01	**0.620** ± 0.04	0.501 ± 0.01	**0.730** ± 0.00	**0.502** ± 0.15
EtL Consensual	0.621 ± 0.09	0.558 ± 0.05	0.499 ± 0.00	0.690 ± 0.04	**0.665** ± 0.00
EtL Non-Consensual	0.604 ± 0.15	0.508 ± 0.18	0.393 ± 0.16	0.629 ± 0.15	**0.554** ± 0.16
AtL Consensual	0.458 ± 0.12	0.429 ± 0.12	0.393 ± 0.02	0.552 ± 0.10	**0.496** ± 0.05
AtL Non-Consensual	0.655 ± 0.04	**0.615** ± 0.05	0.490 ± 0.01	0.693 ± 0.03	**0.499** ± 0.10
HtL Consensual	0.412 ± 0.18	0.476 ± 0.04	0.373 ± 0.03	0.465 ± 0.16	**0.592** ± 0.06
HtL Non-Consensual	0.576 ± 0.18	**0.655** ± 0.03	0.510 ± 0.01	0.642 ± 0.13	**0.538** ± 0.02
w/o EtL Consensual & HtL Non-Consensual & AtL Non-Consensual	0.610 ± 0.16	**0.633** ± 0.05	0.459 ± 0.03	0.655 ± 0.13	**0.609** ± 0.05
w/o EtL Non-Consensual & HtL Consensual & AtL Consensual	0.601 ± 0.16	0.549 ± 0.12	0.460 ± 0.07	0.612 ± 0.20	**0.535** ± 0.13
w/o HtL Non-Consensual	**0.704** ± 0.01	**0.642** ± 0.02	0.498 ± 0.02	**0.728** ± 0.01	**0.631** ± 0.03
w/o AtL Non-Consensual	**0.695** ± 0.02	**0.604** ± 0.01	0.484 ± 0.01	**0.719** ± 0.02	**0.566** ± 0.09
w/o HtL Non-Consensual & AtL Non-Consensual	0.583 ± 0.18	0.578 ± 0.06	0.480 ± 0.02	0.633 ± 0.15	**0.507** ± 0.09
w/o EtL Non-Consensual & HtL Non-Consensual	0.598 ± 0.16	0.534 ± 0.03	0.440 ± 0.02	0.594 ± 0.24	**0.587** ± 0.04
w/o EtL Non-Consensual & AtL Non-Consensual	**0.688** ± 0.04	0.558 ± 0.16	0.476 ± 0.08	**0.715** ± 0.03	**0.545** ± 0.12
w/o EtL Non-Consensual & HtL Non-Consensual & AtL Non-Consensual	**0.688** ± 0.04	**0.619** ± 0.05	0.490 ± 0.02	**0.722** ± 0.02	**0.458** ± 0.08

Table 6: Performance Comparison of BERT uncased Model Trained on the OLID Dataset Across 19 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.

Condition	SBIC (ID)	DynaHate (OOD)	ETHOS (OOD)	OLID (OOD)	ToxiGen (OOD)
*Baseline (100% train)	0.801 ± 0.01	0.634 ± 0.06	0.630 ± 0.03	0.438 ± 0.10	0.529 ± 0.09
w/o EtL Non-Consensual	0.798 ± 0.01	**0.674** ± 0.02	**0.632** ± 0.01	**0.498** ± 0.03	**0.561** ± 0.06
w/o EtL Non-Consensual & EtL Consensual 75%	0.795 ± 0.01	**0.678** ± 0.03	0.630 ± 0.01	**0.512** ± 0.02	**0.620** ± 0.04
w/o EtL Non-Consensual & EtL Consensual 50%	0.795 ± 0.02	**0.695** ± 0.01	**0.631** ± 0.02	**0.511** ± 0.02	**0.534** ± 0.10
w/o EtL Non-Consensual & EtL Consensual 25%	0.789 ± 0.02	**0.695** ± 0.01	0.624 ± 0.02	**0.492** ± 0.04	**0.617** ± 0.06
EtL Consensual	0.779 ± 0.02	**0.674** ± 0.02	0.622 ± 0.01	**0.477** ± 0.03	0.452 ± 0.17
EtL Non-Consensual	0.739 ± 0.03	0.628 ± 0.07	0.587 ± 0.02	0.435 ± 0.05	0.484 ± 0.12
AtL Consensual	0.732 ± 0.00	**0.707** ± 0.00	0.606 ± 0.00	**0.490** ± 0.01	**0.637** ± 0.04
AtL Non-Consensual	0.684 ± 0.10	0.577 ± 0.08	0.560 ± 0.09	0.385 ± 0.10	0.471 ± 0.09
HtL Consensual	0.427 ± 0.11	0.504 ± 0.11	0.370 ± 0.08	0.396 ± 0.11	0.447 ± 0.10
HtL Non-Consensual	0.698 ± 0.04	**0.682** ± 0.03	0.577 ± 0.03	**0.476** ± 0.03	**0.561** ± 0.12
w/o EtL Consensual & HtL Non-Consensual & AtL Non-Consensual	0.742 ± 0.01	0.623 ± 0.04	0.586 ± 0.01	**0.479** ± 0.02	0.506 ± 0.06
w/o EtL Non-Consensual & HtL Consensual & AtL Consensual	0.765 ± 0.05	**0.658** ± 0.07	0.601 ± 0.06	0.391 ± 0.10	**0.545** ± 0.09
w/o HtL Non-Consensual	0.792 ± 0.02	**0.678** ± 0.01	**0.633** ± 0.02	**0.465** ± 0.06	0.482 ± 0.10
w/o AtL Non-Consensual	0.764 ± 0.06	0.616 ± 0.15	0.583 ± 0.09	**0.452** ± 0.11	0.515 ± 0.10
w/o HtL Non-Consensual & AtL Non-Consensual	0.783 ± 0.02	**0.657** ± 0.05	0.628 ± 0.01	0.409 ± 0.06	0.437 ± 0.07
w/o EtL Non-Consensual & HtL Non-Consensual	0.801 ± 0.01	**0.663** ± 0.04	**0.635** ± 0.01	**0.483** ± 0.06	0.504 ± 0.05
w/o EtL Non-Consensual & AtL Non-Consensual	0.784 ± 0.02	**0.697** ± 0.01	0.620 ± 0.01	**0.469** ± 0.09	**0.530** ± 0.09
w/o EtL Non-Consensual & HtL Non-Consensual & AtL Non-Consensual	0.796 ± 0.01	**0.666** ± 0.02	**0.631** ± 0.01	**0.444** ± 0.05	0.443 ± 0.06

Table 7: Performance Comparison of BERT uncased Model Trained on the SBIC Dataset Across 19 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.

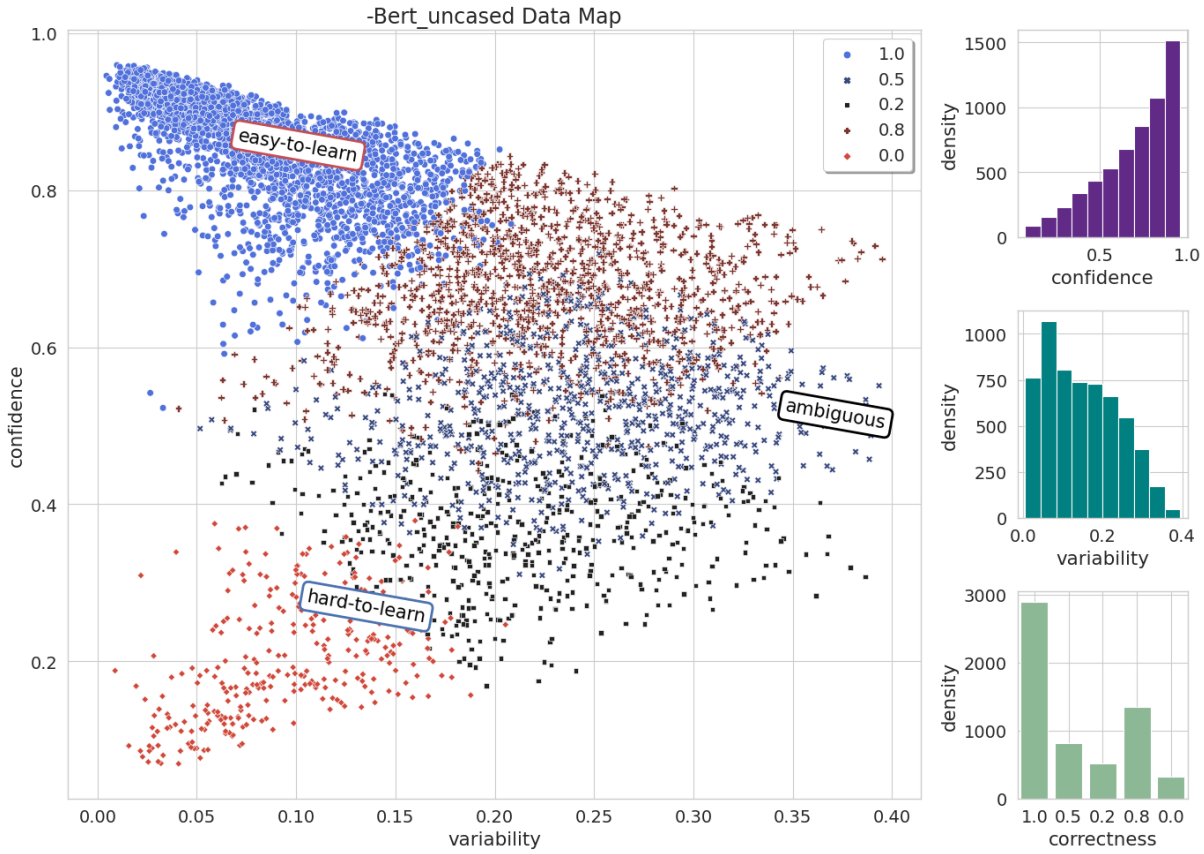

Figure 8: Hard-to-Learn Consensual Datamap

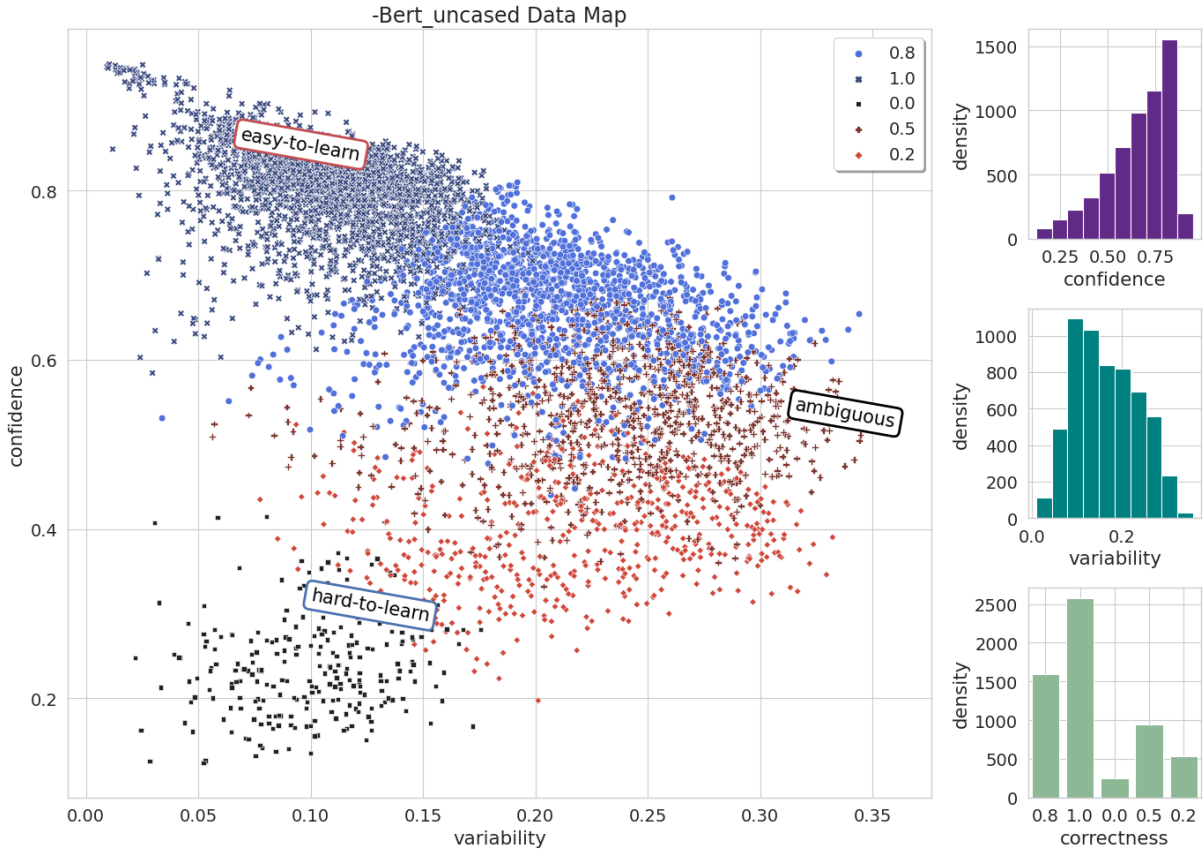

Figure 9: Hard-to-Learn Non-Consensual Datamap

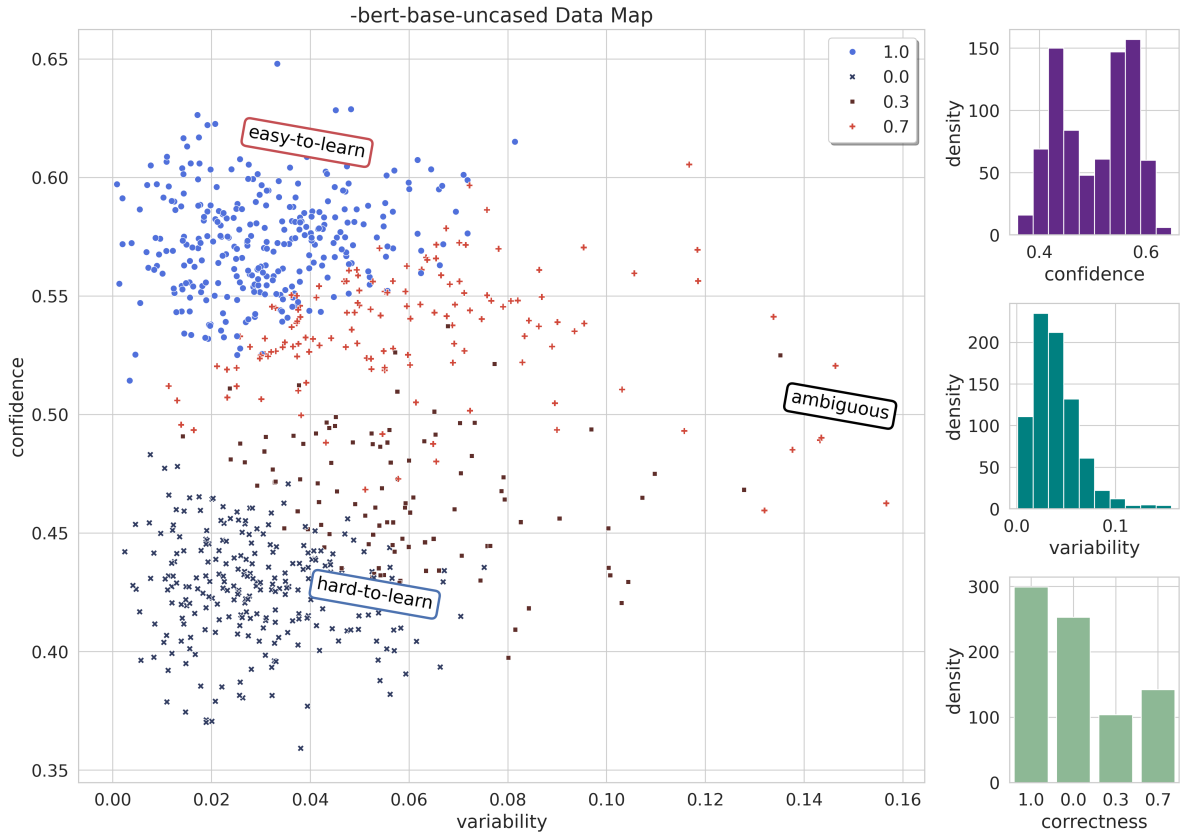

Figure 10: ETHOS Datamap

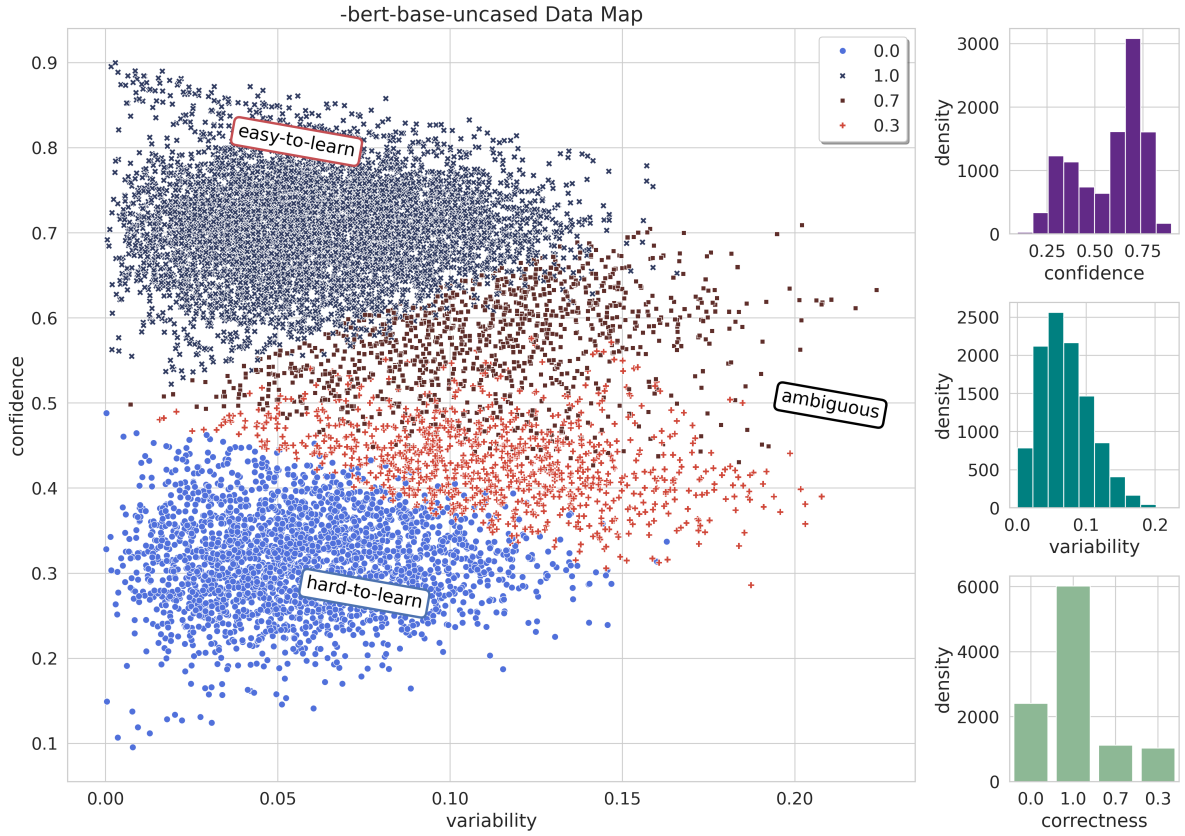

Figure 11: OLID Datamap

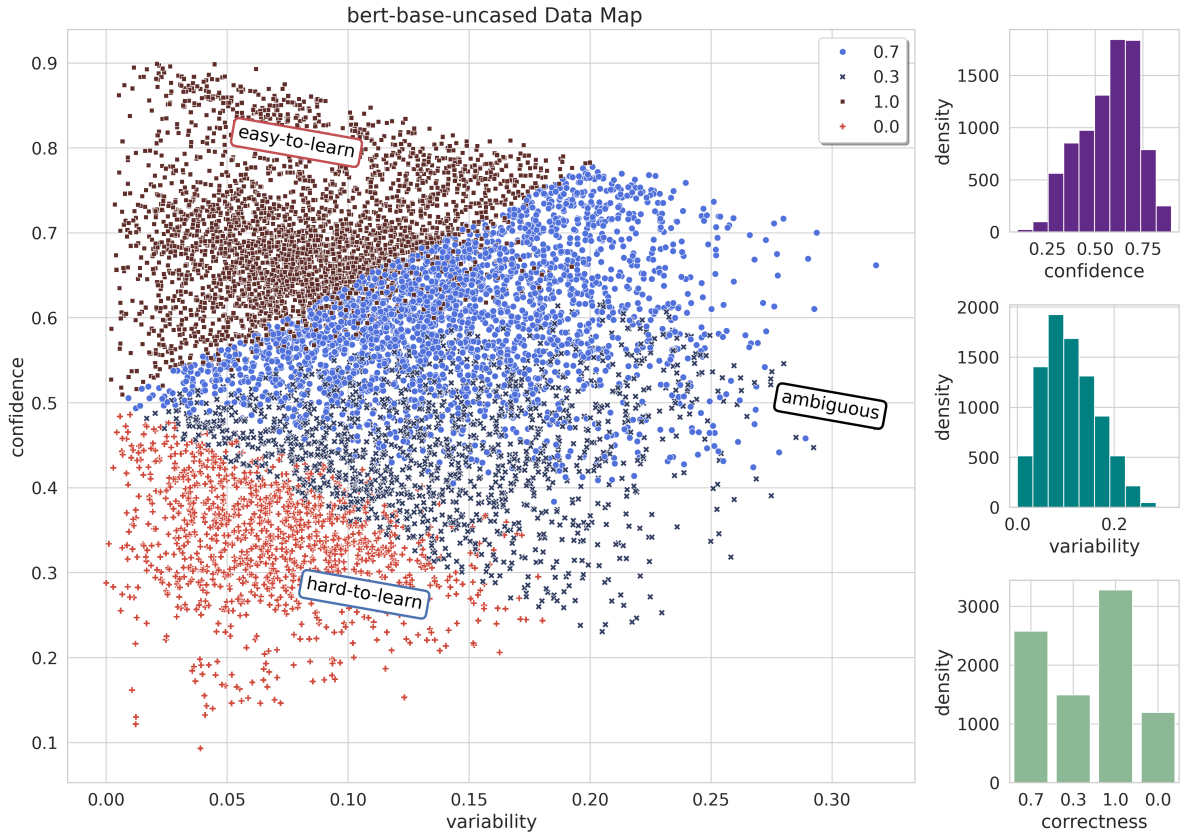

Figure 12: SST-2 Datamap