Smarter, Not Bigger: Unveiling Implicit Hate Speech with Scalable Language Models

Anonymous submission

Abstract

The detection of implicit hate speech is one of the critical challenges faced by the Natural Language Processing community as it requires the use of indirect and vague language-a step that traditional approaches find impossible. An exhaustive study of implicit hate detection has been carried out in this paper, concentrating on a subclass of tasks known as Implicit Target Span Identification, aimed at identifying spans of text that perform less explicit targeting of protected groups.

We systematically evaluate three Masked Language Models—BERT, RoBERTa, and HateBERT—alongside two Small Language Models ModernBERT and SmolLM2 and, as well as Large Language Models with LLama 3.2B and GPT-3.5. Our approach considers both zero-shot and fine-tuning methodologies while examining the effects of instruction tuning and Low-Rank Adaptation (LoRA) to assess their impact on detection tasks.

The results indicate that ModernBERT with only 149M parameters outperforms instruction-tuned larger models such as LLaM 3.2B, ModernBERT achieved F1 scores of 72.2 and 75.1 on IHC and SBIC datasets, respectively, while LLaM 3.2B attained 70.8 and 74.2 F1 scores for IHC and SBIC, respectively. RoBERTa-Large remains the best overall, scoring 72.5 F1 on the IHC dataset and 75.8 on the SBIC dataset. Compared to it, SmolLM2-135M attained an F1 score of 69.0 on IHC and 71.5 on SBIC, still showing competitive performance notwithstanding its smaller size.

1 Introduction

Warning: *This paper contains offensive content and may be distressing.*

Implicit hate speech represents a sophisticated manifestation of prejudice, characterized by the avoidance of overtly offensive language, while still conveying harmful intentions. In contrast to explicit hate speech, which can be recognized through evident lexical indicators, the identification of implicit hate speech necessitates a nuanced comprehension of contextual nuances, cultural references, and concealed meanings. There are several subtler forms of discriminatory language that can arise, including sarcasm, stereotypes, coded language, and doublespeak. For example, the expression "they don't belong here" conveys a sense of exclusion or hostility without explicitly naming a specific group. Likewise, the remark "We should not lower our standards in order to hire more women" can subtly perpetuate gender stereotypes while maintaining an appearance of plausible deniability.

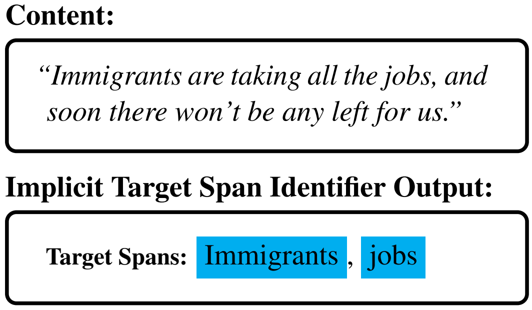

Figure 1: Implicit Target Span Identification Example

The inherent vagueness of implicit hate speech presents a serious challenge to systems of content moderation. As there has been improvement in detection of explicit hate speech, existing models often fail to detect more sophisticated and more covert modes of speech that elude regular detection methods.This limitation highlights the necessity for continued research and the formulation of more sophisticated approaches to effectively address the challenges presented by implicit hate speech.

The detection of implicit hate speech is complicated by its context-dependent nature and dynamic expressions. The primary challenges can be categorized as follows:

Implicit hate speech relies, in many cases, on certain cultural-historical-social context underpinnings. Because of this, it becomes especially difficult to detect it. A single statement may be innocuous in one place and yet invective in another place. This requires models to go beyond mere textual analysis into general context-based knowledge (Gao et al.; Jafari et al.). Such models are important since they must pick up on subtle references and implicit biases contained within everyday language.

Constructing annotated datasets to detect implicit hate speech is a time-consuming and resource-intensive endeavor. It calls for annotators with extensive cultural knowledge, making it challenging and time-consuming (Almohaimeed et al.). Guaranteeing agreement among annotators, especially when working with subjective content, presents an additional layer of challenge. Maintaining the privacy and confidentiality of participants when gathering data from a variety of platforms is another challenge of constructing datasets, albeit critical in training accurate and fair models (Ahn et al.; Ocampo et al., 2023).

Hate speech is an ever-evolving form of expression that, with the use of coded language or dog whistles, is designed as to evade detection systems. Such insidious forms of expression require ongoing development of detection systems, including adversarial learning methods to stay abreast of changing trends of hate speech patterns (Ocampo et al.; Hindy et al., 2022). Models need to be fluid and responsive so that even the most insidious forms of hate speech can be quickly and effectively detected and dealt with.

In order to effectively confront these multifaceted challenges, there exists an imperative for the implementation of innovative strategies in dataset development, an enhanced comprehension of contextual variables, and the formulation of adaptable model architectures. These elements are essential for addressing the complexities inherent in implicit hate speech.

A key element of enhancing detection is the investigation of both Masked Language Models (MLMs) and Large Language Models (LLMs) since they provide synergistic advantages in natural language comprehension (Subramanian et al., 2023). Specifically, MLMs like BERT and HateBERT excel at capturing local linguistic structures and implicit signal pickup at the token level because of their masked token prediction training method. But they find it challenging to concentrate on long-range dependencies and dense context, which are required for detecting implicit hate speech. By comparison, LLMs such as LLaMA 3.2 and GPT-3.5 employ large-scale pretraining on big corpora to better grasp global context and more subtle shades of meaning. Their advanced contextual representations enable them to better identify implicit hostility, especially where hate speech is coded by masked language or cultural references.

By bridging MLMs and LLMs, this work aims to enhance the robustness and interpretability of implicit hate speech detection. MLMs provide a solid foundation for detecting token-level spans, while LLMs introduce knowledge about contextual and dynamic hate speech patterns. LoRA fine-tuning and data augmentation through GPT-based span annotation make the models more adaptable. This joint methodology is significant in meeting the difficulties of identifying implicit hate speech, just as facilitating more precise and equitable content moderation.

This study seeks to fill critical research gaps by posing the following research questions (RQs):

- **RQ1: Does increasing LLM parameter size improve performance on implicit content detection tasks?**

  Larger language models generally perform well for NLP tasks, yet is scale the sole consideration for implicit hate speech detection? Our study examines if larger numbers of parameters lead to meaningful improvements, comparing performance of smaller models (e.g., ModernBERT, SmolLM2) and larger models (e.g., LLaMA 3.2 1B).

- **RQ2: How do instructed LLMs compare to non-instructed LLMs in implicit hate speech detection?**

  Instruction-based fine-tuning has been recognized as an effective method for both model interpretability and task flexibility.

  We look into whether instructed models such as LLama 3.2 1B Instruct outperform their non-instructed versions in implicit hate speech detection.

- **RQ3: What is the comparative effectiveness of few-shot fine-tuning versus full-dataset fine-tuning in the domain of hate speech detection?** Full fine-tuning is computationally costly, requiring substantial amounts of labeled data and computing resources.

Few-shot learning offers a promising alternative, employing a few examples to push the model to make accurate predictions.

Our study examines if few-shot fine-tuning can be as effective as full fine-tuning, particularly in hate speech detection where data accessibility is a main bottleneck.

- **RQ4: Can error analysis of incorrect predictions enhance both explainability as well as model performance?** The implicit hate speech detection continues to be a difficult task due to the shared vocabulary between neutral and hateful speech.

  Through systematic analysis of these errors, we reveal model prediction biases and suggest focused improvements for future fine-tuning approaches.

To address the research questions above, our work makes the following significant contributions:

- **Model Scaling Study in Implicit Hate Speech Detection:** We investigate the effects of scaling up model sizes, showing that larger size does not necessarily equate to better performance—architectural enhancements and domain-specific pretraining are essential.
- **Exploring the Potential of Instruction-Tuned LLMs:** This work conducts a comprehensive comparison of instructed and non-instructed LLMs for implicit hate speech detection, demonstrating the benefits of explicit task conditioning.
- **Few-Shot vs. Full Fine-Tuning:** We examine the effectiveness of few-shot fine-tuning and demonstrate that it is competitive with full fine-tuning, presenting a suitable option when resources are limited.
- **Systematic Error Analysis Using LDA:** We use topic modeling techniques on misclassified predictions, revealing common linguistic patterns that result in false positives and false negatives, and suggesting areas of improvement for future model training.
- **Model Benchmarking on Varied Datasets:** We benchmark a number of models on a combination of the SBIC, IHC, and OffensiveLang datasets, tackling issues of cross-domain generalization and dataset annotation discrepancies.

Our results emphasize the trade-offs between model size, fine-tuning approaches, and instruction tuning, thereby offering novel insights in the new area of implicit hate speech detection. By connecting token-level span annotation and sentence-level classification, we seek to enhance both the interpretability and efficiency of detecting harmful online discourse.

2 Related Work

The field of hate speech detection has undergone remarkable progress, transitioning from conventional machine learning approaches to more sophisticated deep learning models. Initial approaches mainly employed algorithms such as Support Vector Machines (SVMs) and Logistic Regression that leveraged manually crafted linguistic features such as n-grams and sentiment features (Raza et al.; Rawat et al.). Nevertheless, these approaches tended to struggle with detecting implicit hate speech because of their inferior capacity for understanding contextual nuances, leading to high false-negative rates (Reghunathan et al.).

The introduction of deep learning witnessed significant advances with Recurrent Neural Networks (RNNs) and Bi-GRUs, which dealt with sequential dependencies more effectively (Kibriya et al.). The breakthrough, however, came with transformer-based models like BERT, RoBERTA, and HateBERT, which employed contextual embeddings to deal with subtle language more effectively (Aminu et al.). Despite these advances, models trained on explicit content for the most part struggled to detect implicit hate speech, and more specialized methods like Implicit Target Span Detection (ITSD) had to be invented (Jafari et al.).

The implementation of Large Language Models (LLMs) such as GPT-3 and LLaMA introduced novel functionality in capturing hate speech's implicit nature. Instruction-tuned models performed better with the utilization of domain-specific prompts (Kim et al.). Experiments such as (Garg et al.) demonstrated that LLMs, when fine-tuned using adversarial training, were able to pick up on implicit hints that other models overlook. Besides, the application of Low-Rank Adaptation (LoRA) in efficient fine-tuning has been found to be advantageous in achieving a balance between model size and performance, especially in tasks demanding precise token-level span identification.

Data augmentation techniques have been another crucial area of study. For instance, (Ocampo et al., 2023) proposed hard negative sampling to make models more robust by exposing them to diverse linguistic patterns. These techniques help address the menace of coded language, where seemingly harmless-sounding phrases have malicious meanings rooted in shared cultural knowledge. GPT models have also been leveraged to annotate datasets like OffensiveLang, making training data more informative for better detection accuracy.

In spite of such advances, implicit hate speech detection is still a persistent challenge because of the evolving nature of coded terms and the contextual sensitivity of offensive content. More specifically, cross-lingual detection is hard because of cultural and linguistic disparities (Zhang et al.).

This research extends the existing work by investigating the combined application of MLMs and LLMs to advance the detection accuracy as well as interpretability. Our work adds to the literature base by suggesting new target span identification approaches, utilizing contextual embeddings alongside token-level annotations in a bid to overcome the age-old difficulties confronting this field.

3 Implicit Target Span Identification

Implicit Target Span Identification (iTSI) plays a critical role in detecting hate speech, particularly where targeted messages are subtle and contingent on contextual information. Compared to explicit hate speech, implicit forms often rely on cultural or social nuances that standard models fail to appreciate.

Let us consider a text sequence as $C = [t_1, t_2, ., t_n]$, where each element t_i constitutes a token and n is the overall length of the sequence. The objective of iTSI is to predict spans of tokens that implicitly or explicitly mention protected groups. The task is to predict a set $S = (s_{s1}, s_{e1}), ., (s_{sk}, s_{ek})$, where the tuple (s_{si}, s_{ei}) indicates the start and end indices of the i-th target span. The prediction above is done through token-level annotation with the BIO (Begin-Inside-Outside) method so that spans can be identified properly.

The problem is defined formally by the function $f : C \to S$, where f aligns the input text with a set of spans referring to implicit or explicit hate speech targets.

To identify implicit targets, we employ both MLMs like BERT, Hate-BERT, and , which excel at capturing local language patterns, whereas LLMs like GPT-3.5 and LLaMA provide a more general contextual understanding needed to identify subtle implicit content. Additionally, SmolLM2 and ModernBERT, a two smaller-scale language models, are evaluated for its efficiency in implicit hate speech detection. The models are fine-tuned using Low-Rank Adaptation (LoRA), which enhances computational efficiency without a loss in performance.

4 Experimental Setup

4.1 Datasets

To ensure a robust evaluation of our task, we construct a diverse dataset by integrating samples from three prominent sources:

- **SBIC (Social Bias Inference Corpus)** (Sap et al., 2020): This dataset consists of 150,000 structured annotations pertaining to social media posts, encompassing over 34,000 implications across approximately 1,000 distinct demographic groups.

- **IHC (Implicit Hate Corpus)** (ElSherief et al., 2021): A corpus for hate speech detection that consists of a total of 22,056 tweets gathered from eminent extremist groups in the United States that consist of 6,346 tweets that exhibit implicit hate speech.

- **OffensiveLang dataset**(Das et al., 2024) contains 8270 texts generated by ChatGPT. 6616 are labeled "offensive" and 1654 "not offensive." Critically, the dataset was annotated by both humans and ChatGPT.

4.2 Models

To benchmark ITSI's performance, we evaluate multiple architectures under both zero-shot and fine-tuned settings:

- **Masked Language Models (MLMs):** BERT-Base, Hate-BERT, RoBERTa-Large.

- **Large Language Models (LLMs):** LLama 3.2B, GPT 3.5

- **Small Language Model:** SmolLM2-135M and ModernBERT

This setup allows for a comparative analysis of traditional transformer-based architectures versus instruction-tuned LLMs, assessing their effectiveness in detecting implicit hate speech.

4.3 Evaluation Metrics

We employ the following evaluation metrics: precision, recall, accuracy, and F1-score, which measure a model's performance in detecting implicit hate speech at the token level. Analyzing content at the token level rather than at the sentence level ensures granular analysis, allowing offensive spans to be accurately identified and providing an overall evaluation of model performance.

5 Results

5.1 Model Comparison

To benchmark the effectiveness of different architectures in a zero-shot setting, we compare their F1 scores on both IHC and SBIC datasets in Table 1.

Model	#	F1 Score (IHC)	F1 Score (SBIC)
BERT-Base	110M	67.0	63.4
Hate-BERT	110M	68.5	69.2
RoBERTa-Large	355M	72.5	75.8
ModernBERT	110M	72.2	75.1
LLama 3.2 1B	1B	70.8	74.2
SmolLM2-135M	135M	69.0	71.5

Table 1: Zero-shot performance of different models with their number of parameters.

From the results, RoBERTa-Large emerges as the best-performing model, surpassing other architectures in both datasets. ModernBERT follows closely, while models like Hate-BERT and LLama 3.2 1B demonstrate competitive performance, particularly in SBIC. These results emphasize the advantage of larger pre-trained transformers in zero-shot settings, highlighting their ability to generalize across different tasks.

5.2 Few-Shot vs Full Dataset Fine-Tuning

Fine-tuning with limited data (few-shot learning) presents an appealing trade-off between performance and resource efficiency. Table 2 illustrates the comparison between few-shot (FS) and full dataset (FD) fine-tuning for the SmolLM2-135M-Instruct model.

Fine-Tuning Type	IHC				SBIC			
	F1	P	R	Acc	F1	P	R	Acc
FD	66.0	68.0	64.2	92.7	69.8	69.0	70.5	94.0
FS	64.0	66.0	62.0	92.2	68.2	67.0	69.0	93.8

Table 2: FS:Few-Shot Fine-Tuning vs FD:Full Dataset Fine-Tuning

5.3 Zero-Shot Generalization Across Datasets

We assessed models in a **zero-shot** setting on a combination of SBIC, IHC, and OffensiveLang.

Model	Params (M)	IHC+SBIC+OffensiveLang (F1)
BERT-Base	110	63.8
Hate-BERT	110	66.1
RoBERTa-Large	355	72.4
ModernBERT	125	68.9
LLama 3.2 1B	1000	71.5
SmolLM2-135M	135	70.1

Table 3: Zero-shot performance of various models on the combined dataset.

Model	IHC (F1)	SBIC (F1)
Mistral-1B-GPTQ	67.5	71.0
LLama 3.2 1B Instruct	**68.5**	**72.5**
SmolLM2-135M-Instruct	66.0	69.8
SmolLM2-360M	65.0	68.5

Table 4: F1 Score Comparison for Instructed LLMs Vs Non-Instructed LLMs

6 Error Analysis

Evaluation of model performance is an eye-opener with different challenges hindering the precise identification of implicit hate speech. A main problem is that multiple models were observed to be making quite a confident incorrect prediction that label a high number of false positives and quite a few true positives. A number of neutral phrases were flagged as hate speech simply because the model was too sensitive towards specific keywords-the very worries about aggressive terms not to mention that of hatred that does not follow a seemingly direct slur sometimes got forced from the classification. The sentences like "Non-binary is a concept for progressive people" were flagged as offensive, while politically charged hate speech such as "F—ing liberals and their policies" received a non-offensive label. This goes on to indicate the difficulty of the model in understanding the ideological dimension and contextual subtleties.

Detection of implicit offense in the text remained another critique that might take the form of rendered commentary. In other words, statements like "Their traditions are outdated" were labeled neutral even while threaded with stereotyping and dismissive tone. Similarly, when it comes to gender biases, "Women are too emotional for politics" rarely ever slipped through the radar as hate speech. This points out that perhaps greater contextual awareness is required so as to help the model detect subtle offenses.

In the model errors that were found using Latent Dirichlet Allocation (LDA), three major clusters of misclassifying were revealed which include political hate speech, gender stereotypes, and implicit sexism. Clustering analysis suggests that models fail to consistently recognize political language from explicit hate speech, making moderation and detection extremely difficult. Particularly, gender-related biases represented a difficult challenge, because much of the hate speech produced came without the use of extremely aggressive language, while doing a damaging job of reinforcing stereotypes.

A deeper investigation on dataset complexity and inter-annotator agreement constructs focus on the subjective task of detecting implicit hate speech. Results demonstrate more complex datasets leading to higher model error, whilst lower inter-annotator agreement scores build a case regarding flimsy labeling of implicit hate speech. An extended text containing the dataset complexity scores and annotation agreement scores is available in the A.11.

The complexity of annotation calls for improvements in dataset labeling schemes and an enhancement in training of the annotator. Divergences in inter-annotator agreement underline the imperative of utilizing external knowledge bases and context-aware embeddings to minimize inconsistencies. Future works need to probe hybrid annotation schemes that bring together human expertise and AI-assisted labeling for enhanced consistency and reliability in implicit hate speech detection.

7 Discussion

RQ1: Does An Increase in LLM Parameter Improve Performance in Implicit Content Detection Tasks?

The results tell us that larger models typically demonstrate better performance, as in RoBERTa-Large and LLama 3.2 1B Instruct outperforming their smaller counterparts. However, despite having many fewer parameters (355M compared to 1B), RoBERTa-Large attains the maximum level of performance. This is to be explained by a number of different reasons.

The importance of task-specific pretraining is of key importance, in that RoBERTa-Large has been heavily pretrained over a range of different textual corpora using advanced masking strategies that support contextual prediction of words.

Moreover, domain adaptation is also a key point of importance. RoBERTa-Large benefits from earlier work that has been adapted to downstream applications such as implicit content detection, whereas LLama 3.2 is more generally oriented.

Parameter efficiency also plays a key role here. RoBERTa-Large, tuned to linguistic structure, effectively maximizes its parameters in undertaking tasks that require a sophisticated understanding of text.

These results remind us that it is not model size that guarantees better performance. Architectural design, pretraining strategies, and task-specific fine-tuning are all key to determining overall efficacy.

RQ 2: How Do Instructed LLMs Compare to Non-Instructed LLMs in Detecting Implicit Targets?

The instructed LLMs always outperform their uninstructed counterparts, presumably owing to a better comprehension of tasks developed during instruction-based learning. This result supports our hypothesis that models given explicit task-oriented instruction would be more effective in identifying implicit hate speech. An in-depth analysis of results in Table 4 reveals that instructed models achieve higher F1 scores in both datasets, with LLama 3.2 1B Instruct having a one or more point lead over Mistral-1B-GPTQ in the F1 score, while the full table with additional metrics is provided in the Appendix. This result highlights the crucial role that explicit task-oriented instruction plays in improving model performance. The improvements observed can be attributed to the explicit task-oriented instruction given throughout learning, enabling models to better recognize the contextual aspects of implicit hate speech. In addition, instructed LLMs gain from demonstrations that not just focus on a particular task but also contextual, hence improving their generalizability across diverse linguistic frameworks. Such models use their learned knowledge to navigate contextual transitions in cases of culturally sensitive or coded speech more proficiently. The combined impact of organized task guidance combined with contextual flexibility makes instructed LLMs more resilient and reliable in implicit hate speech detection, a notion that is backed up by their higher recall rates and better accuracy.

RQ3: How Effective is Few-Shot Fine-Tuning Compared to Full Dataset Fine-Tuning in Hate Speech Detection?

The findings mostly confirm our hypothesis. Although full-dataset fine-tuning consistently outperforms few-shot fine-tuning, the F1-score difference is comparatively low (≤ 2 points), indicating that few-shot learning is an acceptable substitute when annotation resources are scarce.

The high precision scores (over 92% on both datasets) also demonstrate that few-shot models still generalize well even with limited training sets. Interestingly, precision and recall scores on both environments are still close to one another, which suggests that few-shot fine-tuning does not introduce extreme biases towards false positives or false negatives.

These results highlight the relevance of few-shot learning to the real-world detection of hate speech, where labeled data is generally scarce. Few-shot learning can be combined with data augmentation and self-training techniques in the future to further bridge the performance gap while saving on annotation costs.

The findings mostly confirm our hypothesis. Although full-dataset fine-tuning consistently outperforms few-shot fine-tuning, the F1-score difference is comparatively low (≤ 2 points), indicating that few-shot learning is an acceptable substitute when annotation resources are scarce.

The high precision scores (over 92% on both datasets) also demonstrate that few-shot models still generalize well even with limited training sets. Interestingly, precision and recall scores on both environments are still close to one another, which suggests that few-shot fine-tuning does not introduce extreme biases towards false positives or false negatives.

These results highlight the relevance of few-shot learning to the real-world detection of hate speech, where labeled data is generally scarce. Few-shot learning can be combined with data augmentation and self-training techniques in the future to further bridge the performance gap while saving on annotation costs.

RQ 4: Can Error Analysis of Mismatched Predictions Enhance Explainability?

We analyzed misclassified phrases using Latent Dirichlet Allocation (LDA). Table 5 presents examples of phrases that were incorrectly classified, revealing key thematic clusters where models struggle. Our findings confirm that misclassifications are due to ambiguity, implicit bias, and model-specific limitations. Quite possibly the most intractable obstacle to hate speech detection is semantic overlap between hateful and neutral speech. Phrases such as "white southern Christian" or "immigration laws" are often context-dependent—appearing sometimes in innocuous conversation, other times encoded with hate. Such ambiguity leads to false positives, particularly when the model lacks sufficient discourse-level context.

LDA Topic Cluster	**Example of Misclassified Phrase**
Racial Tension	“white southern christian”
Political Bias	“jewish privilege”
Immigration Debate	“immigration laws”
Conspiracy Theories	“white genocide”
Social Justice	“angry white bigots”
War and Nationalism	“another war for Israel”

Table 5: Examples of Misclassified Topics from LDA Analysis

Inaccurately classified hate speech tends to center around sociopolitical discussion, particularly regarding race, religion, and nationalism. This finding suggests that current models struggle to differentiate between subjectivity and explicit hate, commonly misinterpreting critical discourse or commentary as hostility. These biases are potentially due to imbalances in the training data, where models are exposed to specific linguistic patterns repetitively but fail to adequately differentiate between tone and intent.

Despite significant advancements in implicit hate speech detection, our analysis identifies several persistent failure patterns that hinder model performance. These challenges highlight areas where further refinement is necessary to enhance both accuracy and robustness.

One of the most critical issues is label confusion, particularly in distinguishing between the beginning of a targeted span (B-SPAN) and the continuation of the span (I-SPAN). This misclassification leads to fragmented span detection, diminishing the coherence and reliability of model annotations. Addressing this challenge requires improved sequence modeling techniques that can better capture contextual dependencies between tokens.

Another persistent challenge is ambiguity in context. Implicit hate speech often relies on cultural, historical, or situational knowledge for accurate interpretation. Developing models that incorporate external knowledge graphs or domain-specific embeddings could significantly improve contextual understanding.

A further limitation observed is overgeneralization. Overfitting to specific linguistic patterns leads to poor adaptability, underscoring the need for more robust training strategies that emphasize generalization rather than memorization.

8 Future Directions

Our Future studies will be geared toward several promising directions in detecting implicit hate speech.

One important way should be through improving contextual understanding through Retrieval-Augmented Generation (RAG), where models would retrieve relevant contextual information from external knowledge bases, thus enabling them to detect certain latent forms of implicit hate speech. Furthermore, utilization of larger LLMs with more parameters than those used in our study—like can significantly enhance contextual comprehension and subtlety detection.

Explainability can be enhanced using attention visualization techniques, such as attention heatmaps and transformer attention flow, to allow researchers and moderators to interpret and trust model predictions better.

Data augmentation remains important. Techniques such as paraphrasing with LLMs, back-translation, and adversarial data generation could expand training datasets, rendering models more robust. Interactive visualisation tools combined with explainable AI methods like SHAP values could help demonstrate further insights into model decisions, whereas cross-lingual training and the development of ethically oriented datasets will secure wider applicability and fairness.

9 Conclusion

Our analysis finds salient emphases in the implicit hate speech identification process and discerns that model size in itself cannot solely account for its performance. Although larger models tend to perform well, other factors like parameter efficiency, domain adaptation, and task-specific pre-training are equally important. Apart from being smaller in size, ModernBert and SmolLM2 perform better than the larger ones because of its targeted optimization.

Instruction-based learning has considerably improved the model performance due to its better understanding of context, thus a promising area we see for future research. Also, Few-shot fine-tuning provides a solid alternative to using full-data training with considerable generalizability coupled with simple loss of performance.

Error analysis shows that the perennial difficulty lies in distinguishing between hateful and neutral statements in sociopolitical contexts. To conquer these challenges will need much better domain adaptation approaches and discourse-aware models. Moving ahead, bringing in external knowledge bases and making sure the annotation schema is more robust will indeed be a huge step toward building fair, robust, and scalable hate speech detection systems.

10 Limitations

One key limitation in our study is the model size, our exploration was limited to models up to 3B parameters. Due to this reason, this work wouldn't fully investigate to what extent the increase in parameter sizes could impact detection performance.

Another limitation is around datasets. Our work, restricted by the use of monolingual and non-cross-cultural datasets, could benefit from a deeper exploration of historical trends across diverse sociopolitical contexts and non-English languages. Our study, constrained by monolingual and culturally homogenous datasets, lacks the breadth to fully capture historical trends and evolving sociopolitical contexts across diverse linguistic landscapes.

References

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

Saad Almohaimeed, Saleh Almohaimeed, and Ladislau Bölöni. Transfer learning and lexicon-based approaches for implicit hate speech detection: A comparative study of human and GPT-4 annotation. In *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pages 142–147. ISSN: 2472-9671.

Enesi Femi Aminu, Ayobami Ekundayo, Shedrack David Sarkibaka, Oluwaseun Adeniyi Ojerinde, and Uchenna Cosmas Ugwuoke. Hate speech detector based on hybridized BERT-attention mechanism and context analyzer. 1(1):17–17.

Amit Das, Mostafa Rahgouy, Dongji Feng, Zheng Zhang, Tathagata Bhattacharya, Nilanjana Raychawdhary, Fatemeh Jamshidi, Vinija Jain, Aman Chadha, Mary J. Sandage, Lauramarie Pope, Gerry V. Dozier, and Cheryl D. Seals. 2024. Offensivelang: A community based implicit offensive language dataset. *IEEE Access*, page 1–1.

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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782. Asian Federation of Natural Language Processing.

Samarth Garg, Vivek Hruday Kavuri, Gargi Shroff, and Rahul Mishra. KTCR: Improving implicit hate detection with knowledge transfer driven concept refinement.

Ali Hindy, Varuni Gupta, and John Ngoi. 2022. Classifying and automatically neutralizing hate speech with deep learning ensembles and dataset ensembles.

Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. Target span detection for implicit harmful content. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '24, pages 117–122. Association for Computing Machinery.

Hareem Kibriya, Ayesha Siddiqa, Wazir Zada Khan, and Muhammad Khurram Khan. Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification. 116:109153.

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

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772. Association for Computational Linguistics.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. Hate speech detection in social media: Techniques, recent trends, and future challenges. 16(2):e1648. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1648.

Muhammad Owais Raza, Areej Fatemah Meghji, Naeem Ahmed Mahoto, Mana Saleh Al Reshan, Hamad Ali Abosaq, Adel Sulaiman, and Asadullah Shaikh. Reading between the lines: Machine learning ensemble and deep learning for implied threat detection in textual data. 17(1):183.

Arun Reghunathan, Saummya Singh, Gunavathi R, and Amala Johnson. Advanced approaches for hate speech detection: A machine and deep learning investigation. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, pages 1–5.

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.

Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, and G. Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121.

Yaosheng Zhang, Tiegang Zhong, Tingjun Yi, and Haoming Li. Domain-enhanced prompt learning for chinese implicit hate speech detection. 12:13773–13782.

A Appendix

A.1 Comparing LoRA, VERA, and DORA

To further evaluate the effectiveness of LoRA, we compare its performance against VERA and DORA, two alternative fine-tuning techniques.

Model	F1 Score (IHC)	F1 Score (SBIC)
VERA	68.8	71.2
DORA	69.2	71.5
LoRA	69.5	73.0

Table 6: Performance comparison of VERA, DORA, and LoRA with LLama 3.2 (r=16).

A.2 Comparing LoRA Ranks

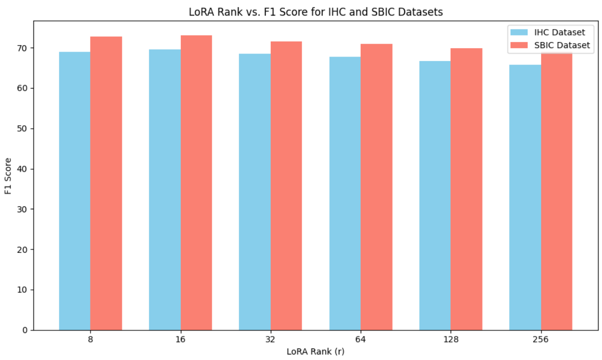

Figure 2: Impact of LoRA rank on F1 scores for IHC and SBIC datasets.

To better visualize the trade-off between computational efficiency and accuracy, Figure 1 below provides a bar chart comparing F1 scores across LoRA ranks for both the IHC and SBIC datasets.

Rank (r)	F1 Score (IHC)	F1 Score (SBIC)
8	69.0	72.8
16	69.5	73.0
32	68.5	71.5
64	67.8	70.9
128	66.7	69.8
256	65.8	68.9

Table 7: Performance of LoRA configurations across datasets.

Table 7 mention that Lower-rank configurations ($r = 8$ and $r = 16$) perform best, balancing computational efficiency and accuracy (Ocampo et al.). Lower-rank configurations ($r = 8$ and $r = 16$) perform best, balancing computational efficiency and accuracy (Ocampo et al.).

The results highlight a key observation: lower-rank configurations ($r = 8$ and $r = 16$) deliver the highest F1 scores while minimizing computational overhead. This suggests that higher-rank values ($r \geq 32$) do not necessarily translate into better performance, potentially introducing unnecessary complexity and resource consumption. These findings align with prior research (Ocampo et al.), reinforcing the idea that smaller, well-optimized LoRA ranks can achieve competitive results without the burden of excessive parameters.

A.3 Comparing LoRA and full-finetuning

Model	Fine-Tuning Type	F1 (IHC)	F1 (SBIC)	Training Time (hrs)
LLama 3.2 1B	Full	70.2	74.5	12
LLama 3.2 1B	LoRA (r=16)	69.5	73.0	3
SmolLM2-135M	Full	68.3	71.0	10
SmolLM2-135M	LoRA (r=16)	67.5	70.2	2

Table 8: Performance and training time comparison between full fine-tuning and LoRA.

A.4 LoRA vs. Full Fine-Tuning

The detailed performance and training time comparison is provided in A.8.

Although full fine-tuning results in slightly higher F1 scores—namely, LLama 3.2 1B from 73.0 to 74.5 on the SBIC benchmark—this minimal gain is at an enormous computational expense. The computational time for full fine-tuning quadruples, from 3 hours using LoRA to 12 hours. This computational cost is even worse for smaller models like SmolLM2-135M, where LoRA is as performant while significantly cutting training time from 10 hours to a mere 2 hours.

A.5 Comparing Instructed LLMs to Non-Instructed

Model	IHC				SBIC			
	F1	P	R	Acc	F1	P	R	Acc
Mistral-1B-GPTQ	67.5	68.5	66.0	92.6	71.0	70.0	71.5	94.0
LLama 3.2 1B Instruct	**68.5**	**69.8**	**67.2**	**93.0**	**72.5**	**71.8**	**73.0**	**94.2**
SmolLM2-135M-Instruct	66.0	68.0	64.2	92.7	69.8	69.0	70.5	94.0
SmolLM2-360M	65.0	67.2	63.5	92.5	68.5	68.0	68.8	93.8

Table 9: Performance Comparison Instructed LLMs Vs Non-Instructed

A.6 ModernBERT Performance on OffensiveLang Dataset

ModernBERT demonstrates a significant leap in performance over traditional models on the OffensiveLang dataset, achieving an impressive F1-score of 0.89. This result highlights its superior capability in identifying implicit hate speech, particularly in challenging contexts where other models struggle.

Model	Precision	Recall	F1-score
TF-IDF + SVM	0.65	0.47	0.55
BERT	0.68	0.54	0.53
DistilBERT	0.71	0.46	0.52
ModernBERT	0.78	1.00	0.89
SmolLM2-135M-Instruct	0.58	0.38	0.46

Table 10: Model performance on the OffensiveLang dataset.

ModernBERT's superior recall rate of 1.00 suggests that it captures a vast majority of offensive content, making it particularly effective in scenarios requiring high sensitivity. In contrast, other models, including DistilBERT and BERT, struggle with recall, indicating difficulty in recognizing nuanced hate speech. The results reinforce the importance of leveraging contextualized embeddings and robust fine-tuning techniques to improve detection accuracy.

Furthermore, an in-depth analysis of annotation agreement across datasets reveals substantial inconsistencies. The complexity of posts in the SBIC, IHC, and OffensiveLang datasets suggests that more contextually rich content poses greater challenges for models, necessitating adaptive training strategies.

A.7 Annotation Agreement

Dataset	Average Complexity Score
SBIC	4.3
IHC	3.9
OffensiveLang	3.6

Table 11: Average complexity of posts across datasets.

Dataset	Agreement Metric	IAA Range
SBIC	Cohen's Kappa	0.65-0.72
IHC	Fleiss' Kappa	0.55-0.60
OffensiveLang	Cohen's Kappa	0.60-0.75

Table 12: Annotation agreement levels across datasets.