AD-ASH An Adaptive Classification Framework for Enhanced Sexual Harassment Detection

Anonymous ACL submission

Abstract

The increasing prevalence of online sexual harassment reports highlights the need for effective automated tools to analyze these personal accounts. In this study, we evaluate a range of models, from neural networks to small and large language models, on the SafeCity dataset to classify incidents of sexual harassment, including commenting, ogling, and groping. We found that different model architectures perform best for different types of harassment, underscoring the need for targeted model selection. Specifically, CNN-RNN models are the most effective for detecting "ogling", BERT-FT excels in identifying "commenting", and DeepSeek7B-FT LLM performs best for "groping" related cases. To integrate these complementary strengths, we introduce AD-ASH, an adaptive ensemble framework that automatically selects the highest-performing model for each category of harassment. By dynamically matching models to task types, AD-ASH achieves state-of-the-art accuracy ranging from 84% to 88% across classes. This adaptive approach offers a robust solution for the nuanced task of harassment classification, demonstrating improved performance over single-model baselines. Our findings highlight the importance of model specialization and ensemble learning in sensitive, real-world applications. Supplementary analyses, including word clustering and LIME-based interpretation of model predictions, are provided in the appendix to offer further insight into language cues that drive classification.

1 Introduction

Social media platforms have significantly affected public discourse by providing spaces in which individuals openly share personal experiences, including sensitive narratives about sexual harassment. Movements such as #MeToo have encouraged countless victims to share their experiences online, creating an extensive, yet linguistically diverse body of narratives. Analyzing these narratives manually is impractical because of their sheer volume and linguistic complexity, necessitating effective automated natural language processing (NLP) tools to classify and detect instances of sexual harassment swiftly and accurately.

Example Instances	C	O	G
"a bunch of guys were passing very bad comments"	1	0	0
"Men and boys hanging around outside the station, staring and passing comments on women passingby."	1	1	0
"a man tried to touch me inappropriately on the road. i looked at him and said what and he didn't react to it. i went away."	0	0	1

Table 1: Annotated example instances from the SafeCity dataset with binary labels for **C** (Commenting), **O** (Ogling), and **G** (Groping). Positive cases (1) are shaded in red, and negative cases (0) are shaded in green.

Early contributions, notably by Karlekar and Bansal (2018), introduced the SafeCity dataset, which comprises approximately 10,000 anonymized victim narratives. Table 1 shows three example narratives from the SafeCity dataset. Initial studies applied neural network architectures such as CNN and RNN to classify harassment types effectively. The emergence of transformer-based models, particularly BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019), has significantly improved contextual representation capabilities and advanced NLP performance considerably.

More recently, large language models (LLMs) such as Llama-3.1 (Grattafiori et al., 2024) and DeepSeek (DeepSeek-AI et al., 2025) have further expanded the toolkit for these tasks. These LLMs support powerful prompting techniques, such as zero-shot, one-shot, and few-shot learning, which have shown notable effectiveness across a range of natural language processing applications. A

study by Brown et al. (2020) showed that GPT-3, using few-shot prompting, can achieve state-of-the-art results on multiple NLP benchmarks by effectively generalizing from a small number of provided examples without requiring additional fine-tuning. Additionally, Retrieval-Augmented Generation (RAG) enables the dynamic retrieval of relevant examples from external sources to guide predictions, thereby enhancing context sensitivity and model flexibility. Lewis et al. (2021) demonstrated that this approach significantly improves performance on knowledge-intensive tasks, particularly in settings with limited supervision or highly variable language, such as those found in personal harassment narratives.

Research has shown that different models, ranging from neural networks to transformer-based SLMs and LLMs, exhibit varying performance across different tasks or subsets of the same task. Studies by Lai et al. (2024); Zhou et al. (2021); Balasubramanian et al. (2018) demonstrated that adaptive or ensemble models can offer benefits by incorporating a diverse set of models, each performing better on specific subsets of a given task.

Building on this insight, our study introduces AD-ASH, an adaptive and extensible ensemble framework that dynamically selects the most effective model for each type of harassment classification task. Our findings reveal that different model architectures specialize in different categories: CNN-RNN performs best for "ogling", BERT-FT for "commenting", and DeepSeek7B-FT LLM for "groping" related cases. This diversity reflects how different narrative types emphasize distinct linguistic structures, making them better suited to different computational approaches. Crucially, we observe that in some cases, LLMs do not consistently outperform smaller, task-focused models (Everitt et al., 2025; Bellos et al., 2024). Simpler models such as CNN-RNN yield more accurate results, particularly in noisy or narrowly defined contexts. These results emphasize that adaptability, not model size, is central to robust classification, and that a one-size-fits-all strategy is insufficient for complex, real-world NLP tasks. Furthermore, the AD-ASH framework is designed with extensibility in mind, allowing new and emerging models or classification techniques to be easily integrated into the adaptive system, ensuring its continued effectiveness as NLP technology evolves.

We also identify dataset noise, such as labeling inconsistencies and ambiguous language, as a significant challenge. These issues can hinder model learning and introduce errors into both training and prediction, disproportionately affecting LLMs that rely heavily on broad contextual generalization (Budnikov et al., 2025). As part of our study, we examine the nature of this noise and its impact on model behavior. Looking ahead, we aim to extend our work by investigating LLM prompting strategies, particularly how dynamic few-shot prompting may help mitigate the effects of dataset misclassifications and improve reliability in noisy, real-world applications.

Our contributions include the following:

1. A comparative evaluation of small and large language models (SLMs and LLMs) in classifying sexual harassment narratives.

2. A systematic exploration of fine-tuning and prompt engineering strategies, ranging from zero-shot and few-shot prompting to Retrieval-Augmented Generation (RAG), to assess their effectiveness in improving classification performance of sexual harassment type detection.

3. The introduction of AD-ASH, a novel adaptive and extensible ensemble framework that dynamically selects the best-performing model for each harassment type, demonstrating improved accuracy across diverse narrative structures.

This paper continues with related work, problem definition, methodology, experimental setup, and detailed results analysis, and concludes with implications and future research directions

2 Related work

Research on analyzing personal narratives of sexual harassment is still developing, with limited work specifically addressing these stories. However, studies in related domains have laid the groundwork for this research. For instance, early efforts to analyze domestic abuse narratives on social media platforms demonstrated the potential of computational methods for extracting valuable insights from sensitive, user-generated content (Schrading, 2015; Schrading et al., 2015). NLP has also been applied to other socially driven tasks, such as abuse detection across social media platforms (Founta et al., 2018) and identifying signs of depression and suicidal ideation in user content (Pestian et al., 2008; Yazdavar et al., 2017).

A key contribution to the analysis of sexual harassment narratives is the SafeCity study by Karlekar and Bansal (2018), which introduced the SafeCity dataset, a large collection of nearly 10,000 victim-reported stories. They applied CNN-RNN architectures to classify harassment narratives into multiple categories, achieving 80-86% accuracy while highlighting the challenge of capturing complex contextual and sequential nuances in these personal accounts.

Recently, transformer-based models, such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) have revolutionized NLP by offering richer contextual representations, leading to significant improvements in various text classification tasks. These models, along with earlier neural networks and more advanced large-language models (LLMs), contribute to a diverse range of models with unique capabilities. Building on these advancements, research has investigated ensemble learning techniques to enhance the performance and leverage the strengths of multiple models within this space. For example, Shahri et al. (2020) combined BERT with CNN and RNN components to capture complementary features and boost the classification accuracy. Similarly, Kim et al. (2021) introduced the auxiliary class-based multiple-choice learning (AMCL) framework, which improves performance through model specialization. Furthermore, Large et al. (2019) showed that combining classifiers from different model families can enhance the predictive accuracy. Drawing on these insights, our study adopts an adaptive ensemble approach for classifying sexual harassment narratives. We expand on the methodology presented by Karlekar and Bansal (2018) by incorporating the top-performing candidate models from binary classifiers to enhance multi-label classification tasks.

Several studies have explored large language models (LLMs) (Paik, 2024; Kwon and Hunjoon Kim, 2024; Riahi Samani et al., 2025) as well as frameworks such as LaMSUM (Chhikara et al., 2025). LaMSUM, a multi-level framework for generating extractive summaries from Safe City posts using LLMs, employs various voting methods for robust summarization. Evaluations of LaMSUM with models such as Llama, Mistral, and GPT-4o highlight its superiority in extractive summarization, highlighting LLMs' strength of LLMs in summarization tasks. However, while LLMs excel at summarization, our findings suggest that they are

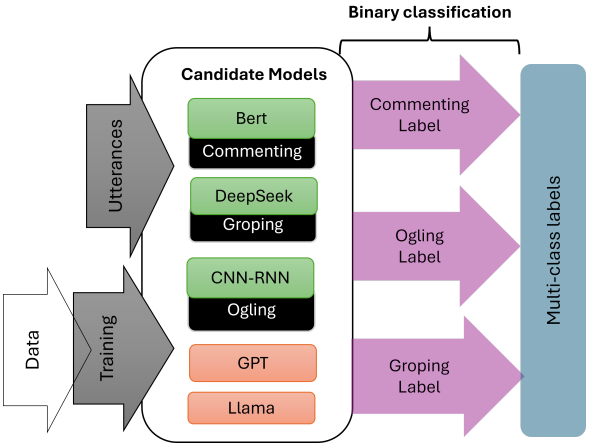


Figure 1: Adaptive model: During the training phase, the candidate models for each type of sexual harassment detection are selected and applied to the binary classification task for each utterance, before being used for the multi-class classification.

less effective in label classification, especially for detecting sexual harassment types. Despite LLMs such as GPT-4o achieving high accuracy in content detection (Wen et al., 2024; Chhikara et al., 2025), our study shows that Small Language Models (SLMs) such as BERT outperform LLMs in accurately classifying harassment labels. This indicates that while LLMs are powerful for summarization tasks, task-specific models, such as SLMs, are more suited to the nuanced task of classifying sexual harassment labels. Moreover, retrieval augmented generation (RAG) methods that dynamically incorporate relevant examples into model predictions have demonstrated superior performance by effectively grounding responses in contextually relevant information (Lewis et al., 2021).

3 Problem definition

In this section, we formally define the problem of detecting sexual harassment. Given an utterance $U = \{w1, ...wN\}$, where w, represents a textual word, the goal is to identify the types of sexual harassment in each utterance from the set $\{Commenting, ogling, groping\}$. The problem is approached as two sub-problems.

The first sub-problem consists of three independent binary classification tasks, each corresponding to one harassment type. In this case, the possible output labels for each classifier are: *[Commenting, Non-Commenting]*, *[Ogling, Non-Ogling]*, and *[Groping, Non-Groping]*. Consequently, three separate binary classifiers are trained, one for each

category.

The second sub-problem is a multi-label classification task, where any combination of the three categories is allowed. This results in $2^3 = 8$ possible label configurations, including a label for none of the three classes.

4 AD-ASH: An Adaptive Architecture for Sexual Harassment Detection

We introduce **AD-ASH**, an adaptive and extensible framework designed for the detection of different types of sexual harassment. The core idea is to evaluate a set of candidate models and select the best-performing one for each harassment category (e.g., commenting, ogling, groping), thereby creating a tailored and modular multi-label classifier.

Figure 1 illustrates the pipeline. During training, each model is assessed for each harassment type in a binary classification setting. The highest-performing model per class is then selected and used in that class inference. The results from the independent harassment type classifiers are combined to produce the final multi-label prediction.

This design not only improves classification accuracy by leveraging the strengths of different models but also ensures extensibility. As new models or fine-tuning techniques emerge, they can be easily integrated into the AD-ASH framework.

4.1 Candidate Models for AD-ASH

Below we describe the models evaluated and used as components within AD-ASH.

BERT-FT utilizes the implementation from the Hugging Face Transformers library "BertForSequenceClassification", which appends a classification head to BERT's final hidden layer to produce logits for each class. During fine-tuning, all model parameters, including the classification head, are optimized jointly. For binary classification, the model outputs two logits passed through a sigmoid activation. For multi-label classification (commenting, ogling, groping), three logits are generated and passed through independent sigmoid functions. Utterances are input directly into the encoder with no additional prompt or instructional text, and gold-standard labels are used for supervision. A maximum sequence length of 512 tokens is used, matching BERT's supported input size.

GPT2-FT leverages the implementation "GPT2ForSequenceClassification", which attaches a fully connected classification head to GPT-2's final hidden representation. All parameters are fine-tuned end-to-end. For binary classification, two logits are produced and passed through a softmax function. For multi-class prediction, three logits are generated and passed through sigmoid activations for each class. Unlike BERT, GPT-2 is guided by short task-specific prompts to steer generation during both training and inference. A maximum sequence length of 512 tokens is used.

In addition to improving classification accuracy, AD-ASH is extensible, new models, prompting strategies, and classification heads can be integrated as the landscape of language models continues to evolve.

Llama-3.1-FT utilizes the Llama-3.1 8B Instruct "LlamaForSequenceClassification" model from the Hugging Face Transformers library to perform binary classification with the "Llama-3.1-8B-Instruct" model. To facilitate binary classification, the model is provided with a task-specific instructional prompt that describes the classification objective clearly (e.g., determining whether a statement reflects commenting, ogling, or groping), enabling the model to align its outputs with the expected label format (e.g., `True`/`False` combinations).

To efficiently manage this large model, we apply 4-bit quantization using the BitsAndBytes library. Additionally, we enhance Llama-3.1 with Low-Rank Adaptation (LoRA) through the PEFT framework. A LoRA configuration is set with a rank of 8, an alpha scaling factor of 16, and a dropout rate of 0.1. To conserve memory during training, gradient checkpointing is enabled and the model is prepared for k-bit training before integrating the LoRA adapters. The fine-tuning process is performed end-to-end on our binary classification task. Input text is tokenized to determine the maximum sequence length and then tokenized with padding accordingly. The tokenized data is converted into PyTorch tensors and structured into a dataset compatible with the Hugging Face Trainer. We optimize the model using cross-entropy loss.

DeepSeek7B-FT We adopt the instruction-tuned "deepseek-llm-7b-chat" model for binary classification using "AutoModelForSequenceClassification". The model is fine-tuned with LoRA and 4-bit quantization similar to Llama-3.1. Instructional prompts are used during training and inference to specify the classification task, enabling the model to produce structured outputs aligned with the required label formats. Tokenization uses `AutoTokenizer`,

and training is performed with the Hugging Face Trainer framework using AdamW optimizer and linear learning rate scheduler.

4.2 Candidate Model Selection for AD-ASH

As previously introduced, AD-ASH is an adaptive and extensible framework that selects the best-performing model for each type of sexual harassment. In our implementation, BERT-FT is selected for detecting "commenting," Deepseek7B-FT for "groping," and CNN-RNN for "ogling", see Figure 1. Each harassment type is first modeled using a binary classifier, and the selected classifiers' outputs are then combined to generate a multi-label prediction for each utterance. This modular design allows the framework to leverage the specific strengths of each model for more accurate and interpretable classification.

5 Experimental setup

5.1 Dataset

In our experiments, we utilize the SafeCity dataset Karlekar and Bansal (2018). The dataset was introduced to capture real-world reports of sexual harassment, SafeCity comprises of 9,892 anonymized narratives where victims describe their experiences along with contextual information such as the incident location. While the original dataset includes annotations for 13 different forms of harassment, our work focuses on a carefully selected subset of categories that are most prevalent in the data: commenting, ogling (staring), and groping (touching). For our experiments, the dataset is partitioned into 7,201 training examples, 990 validation examples, and 1,701 test examples. Given the inconsistent performance across similar models, we investigated potential label noise within the dataset by conducting a validation study. This involved comparing the original dataset labels with expert-provided annotations and applying statistical tests to evaluate mismatch rates and potential directional biases (see Appendix A).

5.2 Evaluation metrics

Single-label. We use accuracy to measure how often a model correctly classifies each sample. Formally, if N is the total number of instances and $\hat{y}_i$ is the predicted label for the i-th instance (with gold label y_i),

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{I}(\hat{y}_i = y_i), \quad (1)$$

where $\mathbf{I}(\cdot)$ is the indicator function that returns 1 if its argument is true, and 0 otherwise.

Multi-label. We report two primary metrics: *exact match ratio* and *Hamming score*. Let each instance have a set of gold labels Y_i (out of L total labels) and a predicted set $\hat{Y}_i$. The exact match ratio is:

$$\text{Exact Match Ratio} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{I}(\hat{Y}_i = Y_i), \quad (2)$$

i.e., the fraction of instances for which the model predicts the exact set of labels. The Hamming score is defined as the complement of the Hamming loss. The Hamming loss is computed as:

$$\text{Hamming Loss} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}_i \triangle Y_i|}{L}, \quad (3)$$

where $\triangle$ denotes the symmetric difference between the predicted labels and the ground truth. The Hamming score is thus,

$$\text{Hamming Score} = 1 - \text{Hamming Loss}. \quad (4)$$

5.3 Baselines

We compare the performance of best performing binary-classification models and our novel adaptive framework for multi-label classification against the following baseline models, as reported by Karlekar and Bansal (2018):

Random forest (Breiman, 2001) is an ensemble learning method that builds multiple decision trees and combines their outputs, thereby improving classification accuracy and reducing overfitting. This traditional approach serves as a useful non-neural reference.

CNN performs sentence classification by transforming input text into word embeddings, then applying multiple convolutional filters with varying kernel sizes to extract local n-gram features. Max-pooling selects the most informative features from each filter, creating a fixed-length representation that feeds into a fully-connected layer for classification probability computation. This approach

processes the entire text as a single instance (Kim, 2014).

RNN utilizes Long Short-Term Memory (LSTM) units (Hochreiter and Schmidhuber, 1997) to capture sequential and contextual dependencies in text. After converting tokenized text into word embeddings, the LSTM network processes them sequentially, maintaining an evolving hidden state that summarizes contextual information. The final hidden state serves as a comprehensive representation for classification through a fully-connected layer. This approach effectively handles tasks where word order and long-range dependencies are crucial.

CNN-RNN combines convolutional and recurrent architectures by first converting text into word embeddings and extracting local features through convolutional filters. These features then feed into an LSTM layer that models temporal dynamics and sequential relationships. The final LSTM hidden state provides a unified representation for classification through a fully-connected layer. This hybrid approach effectively handles scenarios requiring both local patterns and global context (Zhou et al., 2015).

CNN-RNN (B+C)* model combines convolutional layers with a bidirectional LSTM and integrates character-level embeddings to capture fine-grained morphological information. This approach leverages the strength of CNNs for extracting local features and the capability of bidirectional LSTMs for modeling contextual dependencies, as demonstrated previously by Ma and Hovy (2016).

5.4 Training and testing

Across all models, a batch size of 8 is used, and optimization is performed using the AdamW optimizer. A fixed random seed is applied to ensure reproducibility. For single-label classification tasks, standard cross-entropy loss is employed, while for multi-label classification, binary cross-entropy with logits loss (i.e., `BCEWithLogitsLoss`) is used.

At test time, any category whose predicted probability exceeds a chosen threshold of 0.5 is marked as positive, allowing for various combinations of commenting, ogling, and groping to be recognized simultaneously.

5.5 Prompting strategies and Retrieval Setup

In addition to supervised fine-tuning, we evaluate instruction-tuned models using prompting and retrieval-augmented strategies:

Model	Comment	Ogle	Grope
Previous models			
CNN	80.9	82.2	86.0
RNN	81.0	82.2	86.2
CNN-RNN	81.6	**84.1**	86.5
SLMs			
BERT-FT (ours)	**83.2**	83.1	88.0
GPT2-FT (ours)	82.1	83.1	87.4
LLMs			
Llama-3.1 FT (ours)	62.0	59.7	56.0
LLama-3.1 zero-shot (ours)	63	56.7	81.4
Llama-3.1 one-shot (ours)	67.5	59.7	83.6
Llama-3.1 few-shot (ours)	63	55.3	84.4
Llama-3.1 + RAG (ours)	76.6	74.4	83.4
DeepSeek7B FT (ours)	81.7	83	**88.7**
DeepSeek7B zero-shot (ours)	72.8	71.1	79.3
DeepSeek7B one-shot (ours)	51.6	78	82.7
DeepSeek7B few-shot (ours)	60.1	55.8	82.0
Deepseek7B+ RAG (ours)	67.7	60.6	55.3

Table 2: Single-label classification (accuracy) results. The best results are shown in bold, and the second-best results are underlined. Performance of traditional models such as Linear Support Vector Machine, Logistic Regression, Gaussian Naive Bayes, and Support Vector Machine can be found in the original study by Karlekar and Bansal (2018).

Zero-shot prompting : An instruction prompt is constructed using the task definition and a harassment narrative. The model is asked to classify the narrative based on this instruction without any examples.

One-shot and few-shot prompting: Prompts are extended to include one or more annotated examples before the test instance. These examples are manually selected to provide representative context and ensure format consistency.

RAG-enhanced few-shot prompting: We implement a dynamic prompting pipeline using Sentence-BERT (MiniLM-L6-v2) (Hossain et al., 2024) for embedding SafeCity training narratives. FAISS (Douze et al., 2024) is used to retrieve the top-k semantically similar examples (with $k = 6$), balancing class distribution. Retrieved examples are injected into the prompt along with their binary labels. The combined context and test instance are passed to the instruction-tuned LLM (Llama-3.1 or DeepSeek7B), which generates the binary label.

This setup enables us to compare different prompting strategies, including static versus

retrieval-based examples, and evaluate whether the dynamic context improves generalization in binary harassment classification.

6 Results and discussions

Table 2 summarizes the accuracy of various models on binary (single-label) classification tasks for sexual harassment detection. The results indicate that different models excel at different classification tasks. Our **BERT-FT** model achieves the highest accuracy on the *commenting* task, with a score of 83.2%, and the second-highest performance on *groping* at 88.0%. The best performance on the *groping* task is obtained by our **DeepSeek7B-FT** LLM model, with an accuracy of 88.7%. **CNN-RNN** achieves the highest accuracy on the *ogling* task at 84.1%, while our **GPT2-FT** model ranks second for both *commenting* and *ogling*, with accuracies of 82.1% and 83.1%, respectively.

These findings reinforce the notion that different models, ranging from traditional neural networks to SLMs and LLMs, perform better on different aspects of the harassment classification problem. This variability highlights the value of an adaptive architecture that can dynamically leverage the strengths of each model. Overall, SLMs demonstrate strong suitability for domain-specific classification tasks such as these. Table

Additionally, Table 2 reveals that large language models (LLMs), such as Llama-3.1, perform significantly worse on these specialized classification tasks despite their scale and recent development. For example, the Llama-3.1 FT model achieves only 62.0% accuracy on *commenting* and 56.0% on *groping*, highlighting a common limitation of LLMs in extracting precise labels from text, even though they often excel in generative tasks. In contrast, DeepSeek7B-FT, a newer instruction-tuned LLM, demonstrates performance much closer to smaller language models (SLMs), achieving 82.0%, 81.6%, and 86.5% accuracy on *commenting*, *ogling*, and *groping*, respectively. This suggests that some of the typical performance gaps between LLMs and SLMs can be bridged with appropriate tuning and design.

Moreover, the table also shows that LLMs operating in zero-shot, one-shot, few-shot, and RAG-enhanced settings generally underperform when compared to their fine-tuned LLM counterparts and SLMs. Nevertheless, retrieval-augmented generation (RAG) provides noticeable performance improvements for Llama on the *commenting* and *ogling* tasks, indicating the potential of such augmentation strategies. These findings emphasize the need for further exploration of advanced prompting and augmentation methods to identify stronger candidate models for inclusion in the adaptive AD-ASH framework.

Model	**Exact Match**	**Hamming Score**
Random Forest	35.0	70.2
CNN	53.7	80.2
RNN	57.1	81.5
CNN-RNN	59.2	82.3
CNN-RNN(B+C)*	62.0	82.5
GPT2(ours)	62.8	84.0
BERT(ours)	64.7	<u>84.5</u>
DeepSeek7B-FT(ours)	<u>64.9</u>	84.46
AD-ASH(ours)	**66.02**	**85.26**

Table 3: Multi-label classification results. The best results are shown in bold, and the second-best results are underlined.

These performance variations across models may be partially explained by inconsistencies in the original training labels, as identified in our label validation study (see Appendix A). Statistical testing revealed moderate disagreement between the original and expert-assigned labels, with elevated mismatch rates in the `commenting` and `ogling` categories. Additionally, a significant directional bias was found in the `groping` category, suggesting that true groping cases were frequently overlooked. These findings highlight the presence of label noise, which may contribute to variability in model behavior, especially in commenting and ogling categories.

Table 3 summarizes the multi-label classification results across various baseline models, transformer-based small language models (SLMs), large language models (LLMs), and our adaptive ensemble approach. Traditional machine learning methods such as Random Forest yield relatively modest performance, with an exact match score of 35.0 and a Hamming score of 70.2. Neural network architectures, including CNNs, RNNs, and the combined CNN-RNN model, demonstrate improved effectiveness, with the CNN-RNN (B+C)* model achieving an exact match of 62.0 and a Hamming score of 82.5.

Among transformer-based SLMs, BERT-FT delivers the strongest results, attaining an exact match score of 64.7 and a Hamming score of 84.5. GPT2-FT also performs well, with exact match and Ham-

ming scores of 62.8 and 84.0, respectively. Our DeepSeek7B-FT LLM model slightly outperforms the SLMs in terms of exact match with a score of 64.9, while achieving a comparable Hamming score of 84.46. Most notably, our adaptive ensemble model, AD-ASH, which integrates predictions from BERT-FT (for *commenting*), DeepSeek7B-FT (for *groping*), and CNN-RNN (for *ogling*), delivers the highest overall performance. It achieves an exact match score of 66.02 and a Hamming score of 85.26, underscoring the value of selectively combining specialized models for each harassment type.

The improved performance of AD-ASH, the adaptive model suggests that leveraging the complementary strengths of the selected candidate models such as BERT-FT, DeepSeek7B-FT and CNN-RNN, is advantageous for multi-label classification in this domain. While BERT-FT, DeepSeek7B-FT, and CNN-RNN individually capture important contextual and sequential information, their combined predictions offer a more robust representation of the multiple harassment types present in a single narrative. This adaptive, extensible ensemble approach effectively addresses some of the weaknesses inherent in each model when used in isolation, resulting in higher overall classification accuracy. These findings suggest that, given the wide range of available models, solutions could benefit from creating frameworks that leverage the strengths of multiple models. It also emphasizes the importance of continuing to explore older models, as they have proven beneficial when combined with more complex models.

7 Conclusion and future work

In this work, we propose an adaptive model that leverages the strengths of transformer-based small language models (SLMs) such as BERT, large language models (LLMs) like DeepSeek, and neural network architectures such as CNN-RNN. Using the foundational *SafeCity* dataset introduced by Karlekar and Bansal (2018), our goal is to advance the automated classification of sexual harassment narratives. By fine-tuning (FT) a range of models, including BERT, GPT-2, Llama-3.1, and DeepSeek7B, we constructed a pool of candidate classifiers capable of identifying distinct harassment categories. Our adaptive, extensible ensemble strategy selects the top-performing model for each label, BERT-FT for *groping*, DeepSeek7B-FT for *commenting*, and CNN-RNN for *ogling*, based on validation performance. This selective integration enables our system to achieve state-of-the-art results across these key categories.

To support transparency and trust in deployment contexts, we incorporate interpretability techniques such as LIME and t-SNE-based word clustering (detailed in the Appendix B), which reveal important linguistic patterns influencing model decisions.

Future work will focus on refining the ensemble strategy and exploring larger or more optimized variants of DeepSeek and other large language models (LLMs), given their promising performance in sensitive classification tasks. We also plan to investigate more sophisticated fine-tuning methods, advanced ensemble mechanisms, and domain adaptation techniques to further improve robustness and generalizability. In addition, we will explore augmentation strategies using knowledge graphs enriched with relational and contextual information to better capture the nuanced semantics present in harassment narratives.

A critical research direction involves addressing the sensitivity of LLMs to label noise, as highlighted in prior work by Khandalkar et al. (2025); Havrilla and Iyer (2024a). Misclassifications introduced by noisy or ambiguous labels can significantly reduce predictive reliability, especially in emotionally charged and socially sensitive domains like harassment detection. As shown in our label validation analysis (see Appendix A), mitigating these issues will be key to realizing the full potential of LLMs in this space.

Finally, we aim to enhance interpretability using more advanced explanation techniques to better understand multi-label predictions and the model's decision pathways. Ultimately, our work contributes to the broader goal of developing accurate, interpretable, and socially responsible NLP systems that support real-world harassment reporting and victim advocacy (Manche et al., 2025).

8 Limitations

Although our adaptive ensemble framework shows promising performance, it has several limitations. Our experiments rely solely on the SafeCity dataset, which, despite its size, may not capture the full diversity of sexual harassment narratives across different platforms, potentially limiting the generalizability of our findings. Additionally, while transformer-based models like BERT and GPT-2

perform robustly, larger models such as Llama-3.1 and DeepSeek sometimes yield inconsistent results when distinguishing between similar harassment categories, indicating that further adaptation or specialized fine-tuning may be required. Moreover, the inherent imbalance in the distribution of harassment categories and the ambiguity in certain narratives can lead to misclassifications, and our reliance on quantitative metrics like accuracy, exact match, and Hamming score may not fully reflect the qualitative aspects of model predictions. Future work should address these challenges by incorporating more diverse datasets, refining model adaptation techniques, and exploring additional interpretability methods to develop more robust and transparent automated systems for sexual harassment classification.

9 Ethics statement

We analyze sensitive sexual harassment narratives from the anonymized SafeCity dataset, strictly for research and in full compliance with ethical guidelines. Our methods include interpretability analyses to help mitigate potential biases and support victim advocacy, recognizing that automated classification is only part of a comprehensive, human-centered approach. We adhere to all institutional and ACL ethical policies and encourage continued research on the ethical challenges of processing sensitive content.

References

Vivek Balasubramanian, Matteo Turilli, Weiming Hu, Matthieu Lefebvre, Wenjie Lei, Ryan Modrak, Guido Cervone, Jeroen Tromp, and Shantenu Jha. 2018. Harnessing the power of many: Extensible toolkit for scalable ensemble applications. In *2018 IEEE international parallel and distributed processing symposium (IPDPS)*, pages 536–545. IEEE.

Filippos Bellos, Yayuan Li, Wuao Liu, and Jason Corso. 2024. Can large language models reason about goal-oriented tasks? In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 24–34.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Mikhail Budnikov, Anna Bykova, and Ivan P Yamshchikov. 2025. Generalization potential of large language models. *Neural Computing and Applications*, 37(4):1973–1997.

Garima Chhikara, Anurag Sharma, V. Gurucharan, Kripabandhu Ghosh, and Abhijnan Chakraborty. 2025. Lamsum: Amplifying voices against harassment through llm guided extractive summarization of user incident reports. *Preprint*, arXiv:2406.15809.

DeepSeek-AI, Daya Guo, Dejian Yang, and Haowei Zhang et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Tom Everitt, Cristina Garbacea, Alexis Bellot, Jonathan Richens, Henry Papadatos, Siméon Campos, and Rohin Shah. 2025. Evaluating the goal-directedness of large language models. *arXiv preprint arXiv:2504.11844*.

Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2018. A unified deep learning architecture for abuse detection. *Preprint*, arXiv:1802.00385.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alex Havrilla and Maia Iyer. 2024a. Understanding the effect of noise in llm training data with algorithmic chains of thought. *arXiv preprint arXiv:2402.04004*.

Alex Havrilla and Maia Iyer. 2024b. Understanding the effect of noise in llm training data with algorithmic chains of thought. *Preprint*, arXiv:2402.04004.

Lingxiao He and Wu Liu. 2020. Guided saliency feature learning for person re-identification in crowded scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 357–373. Springer.

S Hochreiter and J Schmidhuber. 1997. Long short-term memory. *Neural Comput*, 9(8):1735–1780.

Md Sajjad Hossain, Ashraful Islam Paran, Symom Hossain Shohan, Jawad Hossain, and Mohammed Moshiul Hoque. 2024. Semanticcuetsync at semeval-2024 task 1: Finetuning sentence transformer to find semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1222–1228.

Sweta Karlekar and Mohit Bansal. 2018. SafeCity: Understanding diverse forms of sexual harassment personal stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811, Brussels, Belgium. Association for Computational Linguistics.

Nikhil Khandalkar, Pavan Yadav, Krishna Shinde, Lokesh B Ramegowda, and Rajarshi Das. 2025. Impact of noise on llm-models performance in abstraction and reasoning corpus (arc) tasks with model temperature considerations. *arXiv preprint arXiv:2504.15903*.

Sihwan Kim, Dae Yon Jung, and Taejang Park. 2021. Auxiliary class based multiple choice learning. *Preprint*, arXiv:2108.02949.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Taeksoo Kwon and Connor Hunjoon Kim. 2024. Efficacy of utilizing large language models to detect public threat posted online. *Advances in Artificial Intelligence and Machine Learning*, 04(04):3125–3134.

Zhixin Lai, Xuesheng Zhang, and Suiyao Chen. 2024. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. *Preprint*, arXiv:2403.13335.

James Large, Jason Lines, and Anthony Bagnall. 2019. A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data Mining and Knowledge Discovery*, 33(6):1674–1709.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *Preprint*, arXiv:1603.01354.

Rahul Manche, Fnu Samaah, Tejaswini Tejaswini, and Praveen Kumar Myakala. 2025. Empowering safe online spaces: Ai in gender violence detection and prevention. *Available at SSRN 5176463*.

Seung Yeon Paik. 2024. Analyzing Large Language Models For Classifying Sexual Harassment Stories With Out-of-Vocabulary Word Substitution.

John Pestian, Pawel Matykiewicz, Jacqueline Grupp-Phelan, Sarah Arszman Lavanier, Jennifer Combs, and Robert Kowatch. 2008. Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 96–97, Columbus, Ohio. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ali Riahi Samani, Tianhao Wang, Kangshuo Li, and Feng Chen. 2025. Large language models with reinforcement learning from human feedback approach for enhancing explainable sexism detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6230–6243, Abu Dhabi, UAE. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

J Nicolas Schrading. 2015. Analyzing domestic abuse using natural language processing on social media data.

Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. 2015. #WhyISTayed, #WhyILeft: Microblogging to make sense of domestic abuse. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1281–1286, Denver, Colorado. Association for Computational Linguistics.

Morteza Pourreza Shahri, Katrina Lyon, Julia Schearer, and Indika Kahanda. 2020. Deeppppred: An ensemble of bert, cnn, and rnn for classifying co-mentions of proteins and phenotypes. *bioRxiv*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ruoyu Wen, Stephanie Elena Crowe, Kunal Gupta, Xinyue Li, Mark Billinghurst, Simon Hoermann, Dwain Allan, Alaeddin Nassani, and Thammathip Piumsomboon. 2024. Large language models for

automatic detection of sensitive topics. *Preprint*, arXiv:2409.00940.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-Supervised approach to monitoring clinical depressive symptoms in social media. *Proc IEEE ACM Int Conf Adv Soc Netw Anal Min*, 2017:1191–1198.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A c-lstm neural network for text classification. *Preprint*, arXiv:1511.08630.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018.

A Label Quality Validation Study

To better understand potential sources of variation in model performance, we conducted a validation study to assess the quality of the original labels used in the training data. This study evaluates the level of agreement between the original annotations and a set of expert-verified labels across three harassment categories: `commenting` (verbal), `ogling` (visual), and `groping` (physical).

A.1 Procedure

From the full dataset of 7,201 harassment reports, we drew a stratified random sample of 200 entries from the training set. Each sample was independently reviewed by domain experts and relabeled to form a gold-standard reference set. For each entry, we compared:

- **Original labels:** The initial labels assigned by an automated or external annotation process.
- **Manual labels:** The revised labels assigned by expert annotators, used as ground truth.

Each category label is binary, indicating presence (1) or absence (0). We computed mismatch rates and applied hypothesis tests to characterize the degree and nature of disagreement.

A.2 Mismatch Rates

We first counted how often the original and expert labels differed. Out of 200 examples:

- **Commenting**: 44 instances had mismatched labels (22.0%)
- **Ogling**: 43 instances had mismatched labels (21.5%)
- **Groping**: 28 instances had mismatched labels (14.0%)

These values suggest that the commenting and ogling labels were less consistently annotated than groping. Higher mismatch rates imply greater label noise, which could affect model learning and evaluation, especially for models that are sensitive to nuanced distinctions.

A.3 Statistical Testing

We applied two statistical methods to assess the significance and directionality of the mismatches.

A.3.1 One-Sided Z-Test for Proportions

This test evaluates whether the observed mismatch rate for each category exceeds a predefined acceptable threshold (p_0), such as 5%, 10%, or 15%. It answers the question: **Are the disagreement rates too high to be considered acceptable noise?**

Hypotheses:

H_0: $p \leq p_0$ (Mismatch rate is acceptable)

H_1: $p > p_0$ (Mismatch rate exceeds threshold)

We evaluated p_0 thresholds from 1% to 30%. The test statistic is:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where $\hat{p}$ is the observed mismatch rate and $n = 200$ is the sample size.

Results:

- **Commenting**: The null hypothesis is rejected for thresholds below 18%, meaning that unless we accept an 18% error margin, the mismatch rate is statistically too high.
- **Ogling**: Similar to commenting, the mismatch rate only becomes acceptable at or above 18%.
- **Groping**: The mismatch rate is lower and acceptable for thresholds above 11%.

These results show that the labeling quality for commenting and ogling may not meet stricter quality standards (e.g., 5–10%), which is important when training models requiring high-fidelity labels.

A.3.2 McNemar's Exact Test for Directional Bias

This test determines whether labeling errors were balanced or skewed in a particular direction—i.e., whether the original labels tended to miss positive

cases (false negatives) or incorrectly labeled negative cases as positive (false positives).

We define:

- b: Number of false negatives (original label = 0, new true label = 1)
- c: Number of false positives (original label = 1, new true label = 0)

Under the null hypothesis (no directional bias), b and c should be roughly equal. We use the binomial test:

$$b \sim \text{Binomial}(b + c, 0.5)$$

Results:

- **Commenting**: $b = 20$, $c = 24$, $p = 0.65 \rightarrow$ No directional bias.
- **Ogling**: $b = 24$, $c = 19$, $p = 0.54 \rightarrow$ No directional bias.
- **Groping**: $b = 22$, $c = 6$, $p = 0.0037 \rightarrow$ Significant directional bias.

In the groping category, the number of false negatives significantly exceeded false positives. This suggests the original annotations consistently failed to identify positive cases of groping, which may have led to under-training on this class.

A.4 Summary

This validation study reveals moderate mismatch rates across all three harassment categories. The analysis suggests that, while the labels are broadly usable, elevated mismatch rates in `commenting` and `ogling`, and an asymmetry in `groping`, may introduce noise or directional bias. Such noise can disproportionately affect large language models, which rely heavily on consistent contextual cues to make accurate predictions. Even small amounts of noise in training or prompting data have been shown to substantially degrade LLM performance (Havrilla and Iyer, 2024b). These insights help contextualize the variability we observe in our own LLM results and highlight the importance of label quality when applying large models to sensitive classification tasks.

B Interpretability Analysis

In this section we provide a range of visualization techniques to analyze our best performing model. Each visualization method takes a unique approach, providing fresh insights or reinforcing existing conclusions. These visualizations enhance our understanding of the model, helping to uncover patterns, identify potential issues, and validate assumptions.

B.1 Word clusters

We selected seed words corresponding to class labels and identified the nearest neighbors of each seed word's vector by reducing the dimensionality of the word embeddings using t-SNE (van der Maaten and Hinton, 2008), as shown in Table 4. This visualization not only confirms that our model has effectively learned meaningful word embeddings but also reveals that each type of sexual harassment is associated with a distinct context. Additionally, it demonstrates that our model, AD-ASH, captures related words and concepts specific to each harassment category. We observe that BERT underperformed for the "ogling" category, while the CNN-RNN model used in our adaptive approach achieved better results. This is reflected in the words extracted from our adaptive model, which more accurately represent this specific harassment categories compared to those from the BERT model.

B.2 Saliency Heat Map

Saliency heatmaps (He and Liu, 2020) highlight which words in an input have the greatest impact on the final classification.

In Figure 2a, the word "laughing" has the most significant influence on the classification, followed by "girls" and "noises". These words lead the model to predict the label "commenting", which matches the true label. This corresponds to a scenario where a group of boys makes remarks and strange noises toward girls—behavior that falls under the "commenting" category of sexual harassment.

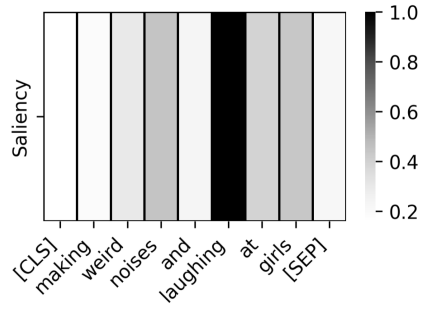
To understand why the model classifies certain incidents as non-commenting, consider Figure 2b. Here, the word "touched", followed by "bus", has the greatest influence, resulting in the model predicting the label "non-commenting", which again aligns with the true label. The model appears to associate "touching" with physical acts such as "groping", which are categorized under a different type of sexual harassment.

B.3 LIME analysis

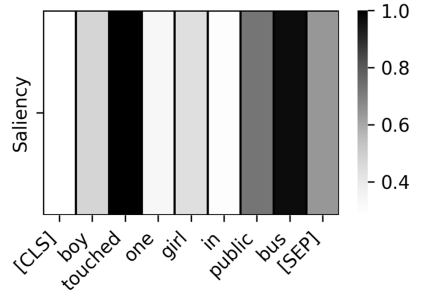
LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) is a technique that

	Observed word clusters				
Model: AD-ASH					
Commenting	shameful	disrespectful	misbehaved	vulgar	inappropriate
Groping	groping	inappropriate touch	assault	harassment	molestation
Ogling	gestures	visually	disturbing	voyeur	leering
Model: BERT					
Ogling	encounter	surrounded	talk	embarrassed	leering

Table 4: Observed word clusters in AD-ASH and BERT.



(a) Correctly classified example of commenting.



(b) Correctly classified example of non-commenting.

Figure 2: Saliency heat-map fo example BERT classified utterances.

helps interpret a model's decision-making process by explaining predictions for specific instances. In the context of our binary classification models, LIME identifies the key features that influence the model's prediction for individual inputs. It does this by approximating the model's decision boundary with a simpler, interpretable model in the local vicinity of the instance, striking a balance between fidelity and interpretability.

This approach provides valuable insights into the features most relevant to a given classification, enhancing our understanding of how the model interprets specific examples. For instance, in the sentence "The guy at first was staring at me and later started passing cheap comments," LIME analysis identified the word "comments" as the most important feature, followed by "passing" and "cheap", indicating the model's recognition of the "commenting" category of sexual harassment. In another example, the phrase "touching/groping, commenting, ogling, and sexual invites" (labeled as "ogling") highlighted the word "ogling" as the most influential feature, demonstrating the model's ability to detect key terms associated with this harassment type. Similarly, in the sentence "A man standing too close to me in a semi-crowded metro station continued to touch me indecently till I pushed him away," LIME identified "touch", "pushed", "standing", and "close" as the most significant terms, aligning with the "groping" classification.

Overall, LIME analysis offers meaningful insights into the linguistic cues driving the model's predictions, contributing to a clearer understanding of how the classifier distinguishes between types of sexual harassment such as "commenting", "groping", and "ogling".