Multi-Agent VLMs Guided Self-Training with PNU Loss for Low-Resource Offensive Content Detection

Han Wang[1], Deyi Ji[2], Junyu Lu[3], Lanyun Zhu[4], Hailong Zhang[2], Haiyang Wu[2], Liqun Liu[2*], Peng Shu[2], Roy Ka-Wei Lee[1*]

[1]Singapore University of Technology and Design
[2]Tencent
[3]Dalian University of Technology
[4]Nanyang Technological University
{han_wang, roy_lee}@sutd.edu.sg, {deyiji, lericzhang, gavinwu, liqunliu, archershu}@tencent.com, dutljy@mail.dlut.edu.cn, lanyun_zhu@ntu.edu.sg

Abstract

Accurate detection of offensive content on social media demands high-quality labeled data; however, such data is often scarce due to the low prevalence of offensive instances and the high cost of manual annotation. To address this low-resource challenge, we propose a self-training framework that leverages abundant unlabeled data through collaborative pseudo-labeling. Starting with a lightweight classifier trained on limited labeled data, our method iteratively assigns pseudo-labels to unlabeled instances with the support of Multi-Agent Vision-Language Models (MA-VLMs). Unlabeled data on which the classifier and MA-VLMs agree are designated as the *Agreed-Unknown* set, while conflicting samples form the *Disagreed-Unknown* set. To enhance label reliability, MA-VLMs simulate dual perspectives, moderator and user, capturing both regulatory and subjective viewpoints. The classifier is optimized using a novel Positive-Negative-Unlabeled (PNU) loss, which jointly exploits labeled, *Agreed-Unknown*, and *Disagreed-Unknown* data while mitigating pseudo-label noise. Experiments on benchmark datasets demonstrate that our framework substantially outperforms baselines under limited supervision and approaches the performance of large-scale models.

Code —
https://github.com/Social-AI-Studio/MA-VLM.git

Introduction

Offensive content on social media, including hate speech, misogyny, and harassment, threatens individual well-being, democratic discourse, and public safety. Although major platforms deploy moderation systems, these often fall short in coverage, fairness, and adaptability across languages, cultures, and modalities. As platforms scale globally and harmful content grows more multimodal and diverse, automated detection becomes essential. Yet building robust and equitable systems remains difficult due to the scarcity of high-quality labeled data, especially for minority languages and underrepresented communities. This scarcity is compounded by the labor-intensive nature of annotation, which demands fine-grained understanding of context, sarcasm, and implicit harm. Consequently, many regions and communities remain underprotected by current AI safety efforts.

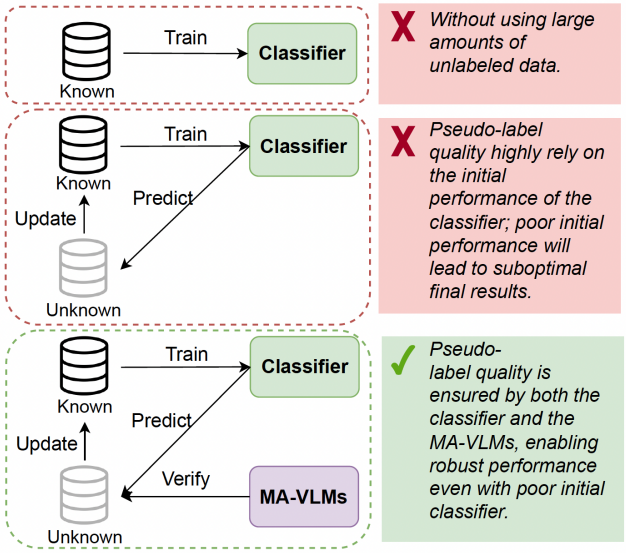

Figure 1: Comparison of our approach (bottom) with supervised-only (top) and traditional self-training (middle).

To tackle offensive content in low-resource settings, researchers have explored several strategies, each with limitations. LLM prompting delivers strong few-shot results but is impractical for large-scale use due to high costs and latency. Data augmentation can help to balance datasets, but is limited to text and often lacks semantic diversity due to extreme label scarcity (Rizos, Hemker, and Schuller 2019; Madukwe, Gao, and Xue 2022; Cao and Lee 2020). Transfer learning from high-resource domains aids generalization but still requires a moderate amount of target-domain labels (Plaza-Del-Arco et al. 2021; Wang, Tan, and Lee 2025; De Oliveira et al. 2024). Self-training bootstraps from unlabeled data but is highly error-prone when initial models are trained with minimal supervision (Alsafari and Sadaoui 2021). Moreover, most methods assume unimodal inputs

*Corresponding authors.

and overlook challenges like label ambiguity and fairness in pseudo-labeling, both critical in real-world application.

We propose a novel self-training framework that addresses these limitations by combining a lightweight classifier with Multi-Agent Vision-Language Models (MA-VLMs) to guide the pseudo-labeling process (Figure 1, bottom). The classifier begins training from a small labeled set and iteratively incorporates high-confidence predictions over unlabeled data. Rather than relying solely on classifier's output, we simulate a multiperspective moderation process where two VLM agents act as a moderator (with safety-first bias) and a user (defending free expression). These agents engage in a prompt-based negotiation on unlabeled samples. If all three agree on a label (positive or negative), the sample is assigned a pseudo-label and added to the training pool as an *Agreed-Unknown* instance. If there is any disagreement between the classifier and either agent, the sample is retained as *Disagreed-Unknown*, a category used to capture uncertain or contested content. This distinction enables the model to treat high-consensus and low-consensus samples differently, enhancing both fairness and robustness in the learning process. This multi-agent negotiation mimics the real-world tension between over-censorship and under-protection, promoting greater fairness and reliability in label propagation.

We also propose a novel Positive-Negative-Unlabeled (PNU) loss to leverage labeled, agreed, and disagreed data. Building on PU learning framework (Elkan and Noto 2008; Sakai et al. 2017), our approach combines: (i) standard positive-negative supervision for labeled data, (ii) soft supervision for *Agreed-Unknown* data to mitigate overfitting to pseudo-labels, and (iii) a PU/NU-style loss for *Disagreed-Unknown* data, modulated by a dataset-specific weighting factor γ. This hybrid formulation enables robust learning under label noise and allows the model to benefit from ambiguous samples often discarded in traditional pipelines.

We evaluate our framework on four benchmark datasets across multimodal and text-only settings, under limited supervision with varying number of labeled examples. Our method outperforms supervised baselines and rivals large VLMs using a lightweight classifier without VLM fine-tuning. Ablation studies validate the contributions of both the MA-VLM prompting format and the PNU loss.

We summarize our key contributions as follows: (i) We propose a novel self-training framework guided by MA-VLMs, designed for scalable and fair offensive content detection in low-resource settings. (ii) We introduce a socially-informed prompting framework that simulates real-world tensions between moderator and user perspectives, enabling more reliable pseudo-label generation. (iii) We propose a PNU-based loss that jointly leverages labeled data, high-confidence pseudo-labels, and disagreement-prone instances via confidence weighting. (iv) We demonstrate strong empirical performance on four benchmark datasets, showing that our method achieves robust results with as few as 50 labeled samples and outperforms larger models in several settings.

Related Works

Low-Resource Offensive Content Detection

Detecting offensive content requires substantial annotated data due to the implicit meanings, cultural context, and subjectivity involved (Lu et al. 2025; Xiao et al. 2024; Hu et al. 2025). However, such data is scarce as offensive instances are often removed by moderation systems, and annotation is costly and complex. In contrast, unlabeled user-generated data is abundant and easily accessible.

When labeled data are scarce, LLMs are a common fallback, leveraging zero- and few-shot capabilities to perform competitively with minimal supervision (Chiu, Collins, and Alexander 2021). Their reasoning also enables explainability via rationale generation (Nirmal et al. 2024; Wang et al. 2023a; Hee and Lee 2025), though high inference costs limit scalability for real-time moderation(Hee et al. 2025; Cao, Lee, and Jiang 2024; Cao et al. 2023).

With moderate supervision, techniques like data augmentation and transfer learning are often used. Augmentation addresses class imbalance through synonym replacement, token warping, or class-conditional generation (Rizos, Hemker, and Schuller 2019; Madukwe, Gao, and Xue 2022; Cao and Lee 2020), but is typically text based and prone to overfitting. Transfer learning adapts models from related domains (Plaza-Del-Arco et al. 2021; Wang, Tan, and Lee 2025; De Oliveira et al. 2024; Wang, Wang, and Lee 2025; Hee, Kumaresan, and Lee 2024), but demands non-trivial labeled data and struggles under extreme label scarcity.

Self-training (Figure 1, middle) is a widely used approach in low-resource settings, where a classifier trained on limited labeled data is iteratively refined using pseudo-labels on unlabeled data. It has shown success in tasks such as text/image classification, NER, and speech recognition (Amini et al. 2025). However, its effectiveness depends on the quality of pseudo-labels. To improve this, methods such as confidence thresholding (Tür, Tür, and Schapire 2005; Lee 2013), proportion-based selection (Zou et al. 2018; Cascante-Bonilla et al. 2021), adaptive thresholding (Wang et al. 2023b; Chen et al. 2023), and dual-classifier frameworks (Karamanolakis et al. 2021; Chen et al. 2021) have been proposed. Some works have applied self-training to VLMs for VQA, image captioning (Yang et al. 2023), and object detection (Zhao et al. 2024; Xu et al. 2023). In offensive content detection, (Alsafari and Sadaoui 2021) uses 9K labeled and 5M unlabeled tweets but struggles in low-resource settings due to pseudo-label bias. Additionally, traditional self-training often overlooks the ambiguity and subjectivity in social tasks, limiting fairness and interpretability.

To overcome the limitations of self-training under scarce labeled data, we draw inspiration from knowledge distillation that transfer general LLM capabilities to task-specific models (Hsieh et al. 2023). Specifically, we propose a self-training framework (Figure 1, bottom) that integrates a lightweight classifier with MA-VLMs. The classifier offers efficient inference, while MA-VLMs validate predictions through a negotiation process. Only samples where both agents and the classifier agree are assigned pseudo-labels and included as Positive or Negative *Agreed-Unknown*. Dis-

agreements are retained as *Disagreed-Unknown*, allowing selective use of uncertain examples during training.

LLMs Based Offensive Content Detection

LLMs are increasingly applied to offensive content detection using prompting strategies such as Zero-Shot, Few-Shot, and Chain-of-Thought (CoT) (Hee et al. 2024). Few-Shot prompting, which incorporates demonstration examples, notably improves GPT-3's performance over Zero-Shot (Chiu, Collins, and Alexander 2021; Wang et al. 2024a). Instruction-tuned LLMs have also enhanced Zero-Shot performance for these tasks (del Arco, Nozza, and Hovy 2023). CoT prompting has been explored to support reasoning in social tasks like hate speech detection (Gupta, Jain, and Bhat 2024; Wang et al. 2023a), though its effectiveness remains limited beyond structured domains such as mathematics (Sprague et al. 2024). Recent studies underscore the potential of multi-agent LLM systems for complex decision-making and simulation (Guo et al. 2024).A related application employs multi-agent debate between LLMs for adversarial attack defense (Chern, Fan, and Liu 2024). However, most existing LLM-based approaches to offensive content detection rely on single models or homogeneous agents, reflecting a narrow viewpoint and neglecting the nuanced trade-off between safety enforcement and free expression.

To this end, we propose MA-VLMs, a multi-agent vision-language framework that simulates diverse social perspectives using two distinct VLMs. Unlike prior single-agent approaches, MA-VLMs introduces a agent-based negotiation mechanism to assess offensive content accounting for both safety enforcement and user expression.

Positive-Unlabeled (PU) Learning

PU learning tackles binary classification using a small set of labeled positives and a large pool of unlabeled data. One strategy applies self-training by initially treating all unlabeled data as negative and iteratively adding confident negatives for retraining (Yu, Zuo, and Peng 2005; Fusilier et al. 2015). Another uses risk estimation; since binary classification requires both positive and negative risks, Elkan and Noto (2008) proposed estimating the negative risk using only positive and unlabeled data, with Kiryo et al. (2017) later introducing a non-negative estimator for greater stability. Traditional PU learning assumes a known positive prior, though some methods estimate it via mixture proportion (Garg et al. 2021). Multi-class extensions (Shu et al. 2020) and recent approaches combining self-training with risk estimation (Chen et al. 2020) have also been proposed.

This concept extends to NU learning, which assumes only negative labels are available. Sakai et al. (2017) further proposed PNU learning, unifying labeled positive, labeled negative, and unlabeled data with a tunable parameter balancing the unlabeled contribution. Although PU learning has been applied in bioinformatics, business analytics, security, and signal processing (Jaskie and Spanias 2019), its use in subjective or social tasks, such as detection of offensive content, where ambiguity and disagreement among annotators are common, remains limited.

To address this gap, we introduce a novel PNU loss inspired by Sakai et al. (2017), adapted for use in our self-training framework. Our loss formulation leverages all available supervision: labeled, confidently pseudo-labeled (*Agreed-Unknown*), and ambiguous (*Disagreed-Unknown*) instances. This design enables effective training in low-resource settings by maximizing the utility of unlabeled data while mitigating the risk of label noise.

Methodology

Detecting offensive content spans diverse subtasks with varying labels, languages, and modalities, reflecting its social complexity and contextual nuance. Some subtasks face a scarcity of labeled data, while unlabeled data is plentiful. This creates a low-resource setting where only a small subset $n \ll N$ of the available N samples is labeled. Our work focuses on leveraging this limited labeled data, with the help of abundant unlabeled data, to enhance detection performance. Figure 2 illustrates the framework employing self-training of a lightweight classifier with pseudo-labels generated by the classifier and MA-VLMs. Classifier retraining employs a PNU loss to integrate all data types. Details on the self-training pipeline, MA-VLMs, and PNU loss follow.

MA-VLMs Guided Self-Training Pipeline

In low-resource settings, classifiers often produce unreliable pseudo-labels, causing error propagation during retraining. To mitigate this, we propose the MA-VLMs guided self-training pipeline (Figure 2), which leverages VLMs' offensive content understanding to verify classifier-generated pseudo-labels. The pipeline consists of five steps:

1. **Train:** The classifier is trained on the n labeled samples.
2. **Predict and Order:** The trained classifier predicts unlabeled samples, ranking them by confidence scores.
3. **Agree or Disagree:** The top k confident predictions are verified by MA-VLMs. Agreed cases are pseudo-labeled as Positive or Negative *Agreed-Unknown*; disagreements remain unlabeled as *Disagreed-Unknown*.
4. **Retrain:** Both agreed and disagreed samples are removed from the unlabeled pool and added to the training set. The classifier is retrained using a novel PNU loss $\mathcal{L}_{pnu}$ that integrates labeled, *Agreed-Unknown*, and *Disagreed-Unknown* samples.
5. **Validation Check:** If development set performance improves after retraining, training advances to the next round; otherwise, the classifier reverts to its prior state, and the current top k pseudo-labels are discarded.

The procedure ends once all unlabeled samples are utilized. By verifying classifier predictions with MA-VLMs and selectively retaining only the top k pseudo-labeled samples that improve performance, our pipeline achieves strong results even with very limited labeled data (e.g., $n = 50$).

Multi-Agent Vision-Language-Models (MA-VLMs)

Current large language model applications in offensive content detection typically adopt a single-agent moderator perspective. In contrast, real-world moderation balances mod-

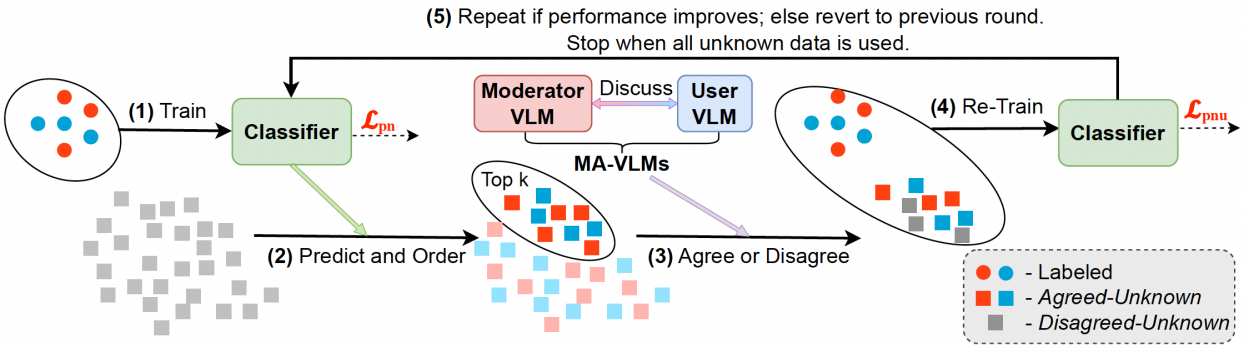

Figure 2: MA-VLMs guided self-training pipeline using PNU loss.

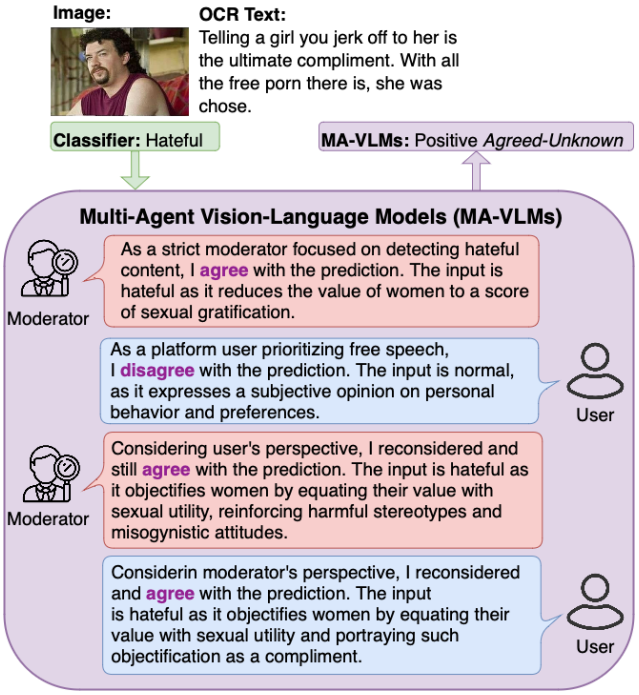

Figure 3: MA-VLMs with hate meme detection example.

erators' focus on safety with users' desire for free expression. To better capture this dynamic, we propose the MA-VLMs prompt format (Figure 3), comprising two VLMs with contrasting personas: a strict moderator and a lenient user. Each agent first provides an initial decision with rationale, then reviews the other's judgment before issuing a final decision. Inputs are labeled *Agreed-Unknown* (Positive/Negative) only if both agents agree with the classifier; otherwise, they are *Disagreed-Unknown*.

In the example shown, a meme initially predicted as hateful by the classifier is first disagreed with by the user agent. After reviewing the rationale of the moderator, the user reviews its judgement, uncovering a nuanced insight previously unrecognized: The meme frames the degradation of women as a compliment, masking its hateful intent. This highlights MA-VLMs' ability to detect implicit hate by combining moderator and user perspectives.

Preliminary of Positive-Unlabeled (PU) Learning

The standard PN learning loss function is:

$$\mathcal{L}_{\text{pn}} = \pi_p \mathcal{L}_p^{y_p} + \pi_n \mathcal{L}_n^{y_n}, \quad \begin{cases} \mathcal{L}_p^{y_p} = \frac{1}{n^p} \sum_{i=1}^{n^p} \ell(g(x_i^p), y_p), \\ \mathcal{L}_n^{y_n} = \frac{1}{n^n} \sum_{i=1}^{n^n} \ell(g(x_i^n), y_n) \end{cases} \tag{1}$$

where π_p and π_n are the positive and negative class priors ($\pi_p + \pi_n = 1$); ℓ is the classification loss; $g(x)$ the model prediction; and y_p, y_n the target labels for positive and negative samples x_i^p, x_i^n, with counts n^p and n^n respectively.

In PU learning, only positive and unlabeled data are available, rendering direct negative loss computation infeasible. To address this, PU learning (Elkan and Noto 2008) approximates negative loss from positive and unlabeled samples. (Kiryo et al. 2017) further propose the non-negative PU loss $\mathcal{L}_{\text{pu}}$ defined below (full derivation in Appendix 1).

$$\mathcal{L}_{\text{pu}} = \pi_p \mathcal{L}_p^{y_p} + \max(0, \mathcal{L}_u^{y_n} - \pi_p \mathcal{L}_p^{y_n}). \tag{2}$$

This loss estimates negatives without negative data and is widely used in fields where positives are easier to obtain, such as bioinformatics, analytics, security, and signal processing (Jaskie and Spanias 2019).

Positive-Negative-Unlabeled (PNU) Loss

Our pipeline uses three data types: labeled, *Agreed-Unknown*, and *Disagreed-Unknown*. *Agreed-Unknown* samples are model-generated without human verification; to mitigate overconfidence, we assign soft targets. *Disagreed-Unknown* data, where the classifier and MA-VLMs disagree, remain unlabeled and are typically discarded in PN learning. Inspired by PNU learning (Sakai et al. 2017), which extends PU learning by incorporating unknown data via a weighting parameter γ, we propose a customized PNU loss to include these samples. The final loss differentiates the three data types to enhance robustness:

$$\mathcal{L}_{\text{pnu}} = \begin{cases} (1-\gamma) \cdot (\mathcal{L}_{\text{pn}} + \mathcal{L}_{\text{soft-pn}}) + \gamma \cdot \mathcal{L}_{\text{pu}}, & \text{if } \gamma \geq 0 \\ (1+\gamma) \cdot (\mathcal{L}_{\text{pn}} + \mathcal{L}_{\text{soft-pn}}) - \gamma \cdot \mathcal{L}_{\text{nu}}, & \text{if } \gamma < 0 \end{cases} \tag{3}$$

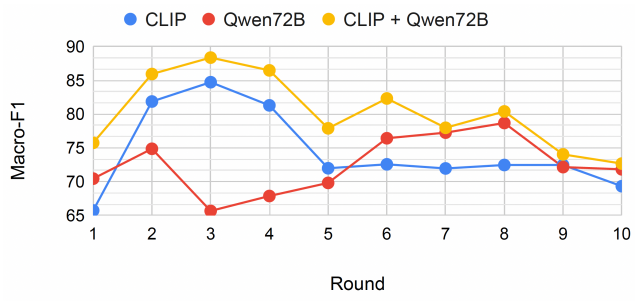

Figure 4: M-F1 of top k pseudo-labeled samples per round on FHM ($n = 100$); only first 10 rounds shown. GT = Ground Truth

where $\gamma \in [-1, 1]$ is dataset-dependent: $\gamma < 0$ yields NU learning (label-reversed PU), $\gamma = 0$ reduces to PN learning, and $\gamma > 0$ corresponds to PU learning. The magnitude of γ controls the strength of PU or NU learning. The PN loss $\mathcal{L}_{\text{pn}}$ follows Equation 1.

The soft-PN loss $\mathcal{L}_{\text{soft-pn}}$ uses *Agreed-Unknown* samples with soft labels to capture uncertainty:

$$\mathcal{L}_{\text{soft-pn}} = \pi_p \cdot \mathcal{L}_{pu}^{\hat{y}_p} + \pi_n \cdot \mathcal{L}_{nu}^{\hat{y}_n} \quad (4)$$

where $\hat{y}_p$ and $\hat{y}_n$ are soft targets for Positive and Negative *Agreed-Unknown* samples x_i^{pu} and x_i^{nu}, with counts n^{pu} and n^{nu}, respectively.

Inspired by PU learning, we treat labeled positives and *Agreed-Unknown* samples as positive data, and *Disagreed-Unknown* samples as unlabeled, yielding the PU loss $\mathcal{L}_{\text{pu}}$:

$$\mathcal{L}_{\text{pu}} = \pi_p(\mathcal{L}_p^{y_p} + \mathcal{L}_{pu}^{\hat{y}_p}) + \max(0, \mathcal{L}_u^{y_n} - \pi_p \cdot (\mathcal{L}_p^{y_n} + \mathcal{L}_{pu}^{\hat{y}_n})) \quad (5)$$

Given the availability of negative data in our setting, the PU loss $\mathcal{L}_{\text{pu}}$ complements the negative component of PN learning, while the NU loss $\mathcal{L}_{\text{nu}}$, formulated similarly, complementing the positive component.

By tailoring losses to each data type, we account for the lower reliability of *Agreed-Unknown* samples and exploit the complementary role of *Disagreed-Unknown* data, leading to a more comprehensive loss formulation.

Experiment

Datasets

To evaluate cross-modal and cross-task generalization, we experiment on four benchmarks: two multimodal and two text datasets. Three target offensive content: *hate speech*, *misogyny*, and *general offensiveness*, while one assesses *sentiment classification*, demonstrating robustness across related tasks. Dataset details follow:

- **Facebook Hateful Memes (FHM)** (Kiela et al. 2020): A multimodal benchmark dataset of 10,000 Facebook memes labeled as *hateful* or *non-hateful*, designed for evaluating hate speech detection in memes.
- **Multimedia Automatic Misogyny Identification (MAMI)** (Fersini et al. 2022): A multimodal dataset of 11,000 Instagram memes annotated for binary classification (Task A: *misogynous* vs. *non-misogynous*) and multi-label classification (Task B: *shaming*, *stereotype*, *objectification*, *violence*).
- **Hate Speech and Offensive Language (HSOL)** (Davidson et al. 2017): A text dataset of 24,783 tweets labeled as *hate speech*, *offensive language*, or *neither*, commonly used in online toxicity detection.
- **Sentiment140 (Sent140)** (Go, Bhayani, and Huang 2009): A large-scale dataset of 1.6 million tweets labeled for *positive* or *negative* sentiment, widely used in sentiment analysis tasks.

We frame all tasks as binary classification. In FHM and Sent140, *hateful* and *negative* samples are treated as Positive labels, while *non-hateful* and *positive* samples are Negative. For MAMI (Task A), *misogynous* is Positive and *non-misogynous* is Negative. In HSOL, *hate speech* and *offensive language* are grouped as Positive labels; *neither* is Negative. All datasets were obtained from Kaggle. The FHM test set lacks labels and is excluded, resulting in 9,000 samples. Each multimodal dataset contains 10,000 samples, while text-only datasets are larger. To ensure balance and manageable training time, we subsample 10,000 balanced examples from HSOL and Sent140 (see Appendix 2 for statistics).

Models and Training Strategies

To assess the effect of our self-training (SelfTrain) pipeline, we firstly compare it against the supervised-only (SupOnly) baseline. In the SupOnly setting, models are trained exclusively on n labeled samples without access to unknown data. Three representative models are evaluated under this setup.

- **Qwen-2.5-VL-7B (Qwen7B)** (Wang et al. 2024b): Alibaba's vision-language model, effective in fine-grained tasks such as VQA and captioning, is well-suited for multimodal offensive detection due to its robust multimodal and human intent understanding.
- **RGCL** (Mei et al. 2023): A retrieval-guided contrastive framework using CLIP's frozen embeddings to align inputs with hateful examples, achieving state-of-the-art hate meme detection.
- **CLIP-Large (CLIP)** (Radford et al. 2021): A contrastive vision-language model projecting images and texts into a shared embedding space, has wide application in multimodal classification.

For SelfTrain (Figure 2), we use CLIP-Large as the classifier and a frozen Qwen-2.5-VL-72B (Qwen72B) (Wang et al. 2024b) as the VLM in MA-VLMs. To assess the benefit of verifying classifier predictions with MA-VLMs, we compare three pseudo-labeling variants: (1) classifier-only, (2) MA-VLMs-only, and (3) their combination.

Experiments Setting

Each dataset is split into 80% training, 10% development, and 10% test sets. Within the training set, only $n = 100$ samples are labeled; the rest remain unlabeled but are used for training. The development set is used for early stopping and the retention or removal of top k pseudo-labels after each round, while the test set is reserved for final evaluation.

Model	#Params	Train	PL Model	FHM		MAMI		HSOL		Sent140	
				Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1
Qwen7B	7B	SupOnly	-	70.78	70.41	**76.20**	**76.06**	85.00	84.89	**78.20**	**78.19**
RGCL	428M	SupOnly	-	35.78	26.35	50.00	33.33	48.00	32.43	48.10	33.28
CLIP	428M	SupOnly	-	64.11	59.24	63.50	62.18	85.30	85.30	64.40	64.22
CLIP	428M	SelfTrain	CLIP	72.11	70.00	68.70	67.03	86.50	86.48	73.10	73.05
CLIP	428M	SelfTrain	Qwen72B	65.33	65.22	69.00	67.42	81.10	81.06	75.70	75.57
CLIP	428M	SelfTrain	CLIP + Qwen72B	**74.22**	**72.68**	73.80	73.49	**86.70**	**86.69**	77.40	77.11

Table 1: Comparison of models and training strategies across four datasets ($n = 100$). PL Model = pseudo-labeling model. Top values highlighted; second-best underlined.

Text	GT	CLIP	Qwen72B	CLIP + Qwen72B
do you know how to use oven! or dig up hitler so he can show you	*hateful*	*hateful*	*hateful*	Positive *Agreed-Unknown*
kermit the frog definitely not a muslim	*hateful*	*hateful*	*non-hateful*	*Disagreed-Unknown*
jerk off to a girl is compliment, with free porn, she was chosen	*non-hateful*	*hateful*	*hateful*	Positive *Agreed-Unknown*

Table 2: Example of pseudo-labeling on FHS dataset; only text is shown for clarity.

Qwen7B is fine-tuned using LoRA. Both RGCL and our method use pretrained CLIP-Large: RGCL extracts features to train a separate classifier, while our model fine-tunes CLIP-Large with a one-layer MLP. All models are trained for 10 epochs, with the best epoch selected based on development performance. Evaluation uses Accuracy (Acc) and Macro-F1 (M-F1), with M-F1 as the primary metric. We set $k = 500$ to balance performance and efficiency. For PNU loss, cross-entropy is used as $\mathcal{L}$, with $y_p = 1$, $y_n = 0$, and soft targets $\hat{y}_p = 0.67$, $\hat{y}_n = 0.33$. The class prior π_p is fixed at 0.5, as deviations introduced class bias and degraded performance (Appendix 3.1). Optimal γ values are determined via ablation: $\gamma = 0.0$ for FHM and $\gamma = 0.1$ for others.

Results

Table 1 shows the performance of three models under SupOnly and three pseudo-labeling variants under SelfTrain, using $n = 100$ labeled samples across four datasets.

SelfTrain CLIP consistently outperforms SupOnly CLIP across all datasets and pseudo-labeling variants, highlighting the value of leveraging unlabeled data. Among the pseudo-labeling variants, the combination of CLIP and Qwen72B achieves the highest performance, surpassing either model individually by over 1.59 M-F1 points, except on HSOL where CLIP-only yields comparable results. These results show that the complementary strengths of CLIP and Qwen72B, combined with disagreement-based filtering, improve pseudo-label quality and overall performance.

Overall, SelfTrain CLIP with CLIP + Qwen72B pseudo-labeling achieves the best results on FHM and HSOL datasets, and ranks second on MAMI and Sent140 datasets, closely trailing the much larger SupOnly Qwen7B while consistently outperforming SupOnly RGCL. SupOnly Qwen7B's strong performance is attributed to its scale and extensive pretraining, though its large size limits its broader application. Although RGCL excels in high-resource settings (Table 3), it struggles in low-resource scenarios due to its reliance on aligning inputs to labeled hateful samples, which hinders convergence under limited labeled data. In contrast, our SelfTrain CLIP with CLIP + Qwen72B pseudo-labeling is more effective in low-resource settings.

Pseudo Label Analysis

The strong performance of SelfTrain CLIP with CLIP + Qwen72B pseudo-labeling is primarily due to large-model verification and disagreement-based filtering, which improve pseudo-label quality. As shown in Figure 4, the CLIP + Qwen72B setup consistently yields higher M-F1 for pseudo-labeled samples, particularly in early rounds. Performance declines as confident samples are exhausted, leaving ambiguous cases. This trend supports our strategy of confidence-based top-k selection to ensure label quality.

Table 2 shows pseudo-labeling examples from the FHM dataset (additional examples in Appendix 4). When CLIP and Qwen72B agree, inputs are typically clearly hateful, even if mislabeled in the ground truth (e.g., the last example), likely mislabeled due to annotator bias, which is a common issue across datasets. Disagreements reflect ambiguity, as in the middle example, where the hateful label is subjective and context-dependent. These observations support our strategy of assigning agreed samples as positive/negative and treating disagreed ones as unlabeled.

Ablation Study

Number of Known Data (n)

To evaluate the robustness of our self-training pipeline across varying labeled data sizes, Table 3 presents the performance of four models: Qwen7B, CLIP (SupOnly), RGCL, and CLIP (SelfTrain) with CLIP + Qwen72B pseudo-labeling, on two multimodal datasets for $n = 50, 100, 250$, and full (all labeled). RGCL results for low-resource settings are omitted due to convergence failure (full results in Appendix 3.2). In extreme low-resource settings ($n = 50$), SupOnly models struggle with scarce labels, whereas SelfTrain maintains M-F1 scores above 70, evidencing its ef-

n	Model	Train	FHM		MAMI	
			Acc	M-F1	Acc	M-F1
50	Qwen7B	SupOlny	64.22	39.11	73.20	73.00
50	CLIP	SupOlny	63.89	48.76	59.50	56.03
50	CLIP	SelfTrain	73.00	71.27	71.40	71.40
100	Qwen7B	SupOlny	70.78	70.41	**76.20**	76.06
100	CLIP	SupOlny	64.11	59.24	63.50	62.18
100	CLIP	SelfTrain	74.22	72.68	73.80	73.49
250	Qwen7B	SupOlny	77.67	75.88	**76.20**	**76.18**
250	CLIP	SupOlny	73.56	69.67	65.70	64.67
250	CLIP	SelfTrain	74.89	72.97	75.00	74.79
full	Qwen7B	SupOlny	**78.67**	**76.00**	71.80	70.20
full	RGCL	SupOlny	78.22	76.17	76.10	75.34
full	CLIP	SupOlny	76.89	75.02	72.00	71.06

Table 3: Comparison of known data sizes (n = 50 to full) on two multimodal datasets. full = all training data used. Top values highlighted; second-best underlined.

PF	Model	Train	FHM		MAMI	
			Acc	M-F1	Acc	M-F1
ZS	Qwen72B	No	74.89	74.46	79.40	79.17
FS	Qwen72B	No	71.22	71.09	75.80	75.08
CoT	Qwen72B	No	**75.33**	74.43	78.30	78.28
MA	Qwen72B	No	75.00	**74.62**	**81.70**	**81.64**

Table 4: Comparison of prompt formats on two multimodal datasets. PF = Prompt Format; ZS = Zero-Shot; FS = Few-Shot; CoT = Chain-of-Thought; MA = MA-VLMs. Top values highlighted; second-best underlined.

fectiveness. In moderately low-resource settings (e.g., $n = 250$), our model underperforms Qwen7B, likely due to Qwen7B's greater capacity for complex tasks. In the high-resource setting, performance of Qwen7B and CLIP on the MAMI dataset unexpectedly declines from $n = 50$ to full, likely due to human annotation errors discussed in the pseudo-label analysis, which limit gains from additional data. RGCL's superior performance in the full setting demonstrates its robustness to these errors.

Prompt Format

Since our pseudo-labeling depends on frozen VLMs (Qwen72B), prompt format selection is crucial. Table 4 reports the performance of four prompting strategies: Zero-Shot, Few-Shot, CoT, and our MA-VLMs, on two multimodal datasets (full results in Appendix 3.3). Few-Shot underperforms Zero-Shot, likely due to bias from randomly selected or unrepresentative examples. CoT offers no improvement, as its step-by-step reasoning is better suited to logical tasks than to social tasks requiring nuanced human judgment. Overall, MA-VLMs consistently outperform other formats, especially on MAMI, likely due to misogyny's more ambiguous definition compared to hate speech. This highlights MA-VLMs' strength in handling label ambiguity through negotiation between agents.

γ	Model	Train	FHM		MAMI	
			Acc	M-F1	Acc	M-F1
- 0.1	CLIP	SelfTrain	69.11	68.35	64.10	59.98
0.0	CLIP	SelfTrain	74.22	**72.68**	70.10	68.84
0.1	CLIP	SelfTrain	73.22	71.50	**73.80**	**73.49**
0.2	CLIP	SelfTrain	**74.44**	71.79	72.60	72.42

Table 5: Comparison of γ values $[-0.1, 0.0, 0.1, 0.2]$ in PNU loss on two multimodal datasets ($n = 100$). Top values highlighted; second-best underlined.

γ Value in PNU Loss

The parameter γ governs the balance between PU/NU and PN learning. Table 5 presents results for $\gamma \in -0.1, 0.0, 0.1, 0.2$ in the PNU loss used in SelfTrain CLIP with CLIP + Qwen72B pseudo-labeling on two multimodal datasets (full results in Appendix 3.4). Performance declines when $\gamma < 0$, indicating NU learning is ineffective in our setting. Similarly, large values ($\gamma > 0.1$) degrade performance, suggesting PU influence should be moderate. Optimal performance occurs for γ between 0.0 and 0.1, depending on dataset characteristics. Offensive content datasets are often constructed using offensive lexicons, causing normal-labeled samples to retain offensive cues. In balanced datasets like MAMI and HSOL, this bias increases the proportion of Positive *Agreed-Unknown* samples, making $\theta = 0.1$ more effective. In contrast, FHM's imbalanced distribution (3:7 positive to negative) results in a more balanced *Agreed-Unknown* set due to this bias, reducing the benefit of higher γ values. For sentiment classification, such bias is absent, yielding comparable performance for $\theta = 0.0$ and 0.1.

Conclusion

In summary, we propose a Multi-Agent Vision-Language Models (MA-VLMs)-guided self-training framework with a tailored PNU loss for low-resource offensive content detection. Our experiments validate the effectiveness of key components: the dual-perspective prompt format (moderator and user) in MA-VLMs, verification of classifier predictions via MA-VLMs, and the PNU loss, which integrates labeled data with *Agreed-Unknown* and *Disagreed-Unknown* pseudo-labels. Moreover, the framework is both task- and model-agnostic, applicable beyond offensive content detection (e.g., sentiment classification), with no restrictions on classifier architecture or VLM inference.

Overall, the lightweight classifier trained via our self-training framework achieves performance comparable to much larger models, demonstrating robustness in extremely low-resource scenarios. This enables scalable, high-quality moderation of social media content in low-resource settings—such as under-resourced languages, modalities, or subtasks—requiring only tens to hundreds labeled samples, and highlights the potential social impact of our approach in promoting safer online environments.

Acknowledgements

This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore, and the Ministry of Digital Development and Information, under its Online Trust and Safety (OTS) Research Programme (Award Grant No. S24T2TS007), and Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 2. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Prime Minister's Office, Singapore, or the Ministry of Digital Development and Information and the Ministry of Education, Singapore.

References

Alsafari, S.; and Sadaoui, S. 2021. Semi-supervised self-training of hate and offensive speech from social media. *Applied Artificial Intelligence*, 35(15): 1621–1645.

Amini, M.-R.; Feofanov, V.; Pauletto, L.; Hadjadj, L.; Devijver, E.; and Maximov, Y. 2025. Self-training: A survey. *Neurocomputing*, 616: 128904.

Cao, R.; Hee, M. S.; Kuek, A.; Chong, W. H.; Lee, R. K.-W.; and Jiang, J. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5244–5252.

Cao, R.; and Lee, R. K.-W. 2020. HateGAN: Adversarial Generative-Based Data Augmentation for Hate Speech Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Cao, R.; Lee, R. K.-W.; and Jiang, J. 2024. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM Web Conference 2024*, 4575–4584.

Cascante-Bonilla, P.; Tan, F.; Qi, Y.; and Ordonez, V. 2021. Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6912–6920.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. SoftMatch: Addressing the Quantity-Quality Tradeoff in Semi-Supervised Learning. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Chen, X.; Chen, W.; Chen, T.; Yuan, Y.; Gong, C.; Chen, K.; and Wang, Z.-H. 2020. Self-PU: Self boosted and calibrated positive-unlabeled training. In *International Conference on Machine Learning*, 1510–1519. PMLR.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2613–2622.

Chern, S.; Fan, Z.; and Liu, A. 2024. Combating adversarial attacks with multi-agent debate. *arXiv preprint arXiv:2401.05998*.

Chiu, K. L.; Collins, A.; and Alexander, R. 2021. Detecting Hate Speech with GPT-3. *arXiv preprint arXiv:2103.12407*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 512–515.

De Oliveira, A. B.; Baptista, C. D. S.; Firmino, A. A.; and De Paiva, A. C. 2024. A Large Language Model Approach to Detect Hate Speech in Political Discourse Using Multiple Language Corpora. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 1461–1468.

del Arco, F. P.; Nozza, D.; and Hovy, D. 2023. Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics.

Elkan, C.; and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 213–220. ACM.

Fersini, E.; Gasparini, F.; Rizzi, G.; Saibene, A.; Chulvi, B.; Rosso, P.; and Sorensen, J. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 533–549.

Fusilier, D. H.; Montes-y Gómez, M.; Rosso, P.; and Cabrera, R. G. 2015. Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management*, 51(4): 433–443.

Garg, S.; Wu, Y.; Smola, A. J.; Balakrishnan, S.; and Lipton, Z. C. 2021. Mixture proportion estimation and PU learning: A modern approach. In *Advances in Neural Information Processing Systems*, volume 34, 8532–8544.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. Technical Report 1(12), CS224N Project Report, Stanford.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; and Zhang, X. 2024. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. *arXiv preprint arXiv:2402.01680*.

Gupta, P.; Jain, N.; and Bhat, A. 2024. Hate Speech Detection using CoT and Post-hoc Explanation through Instruction-based Fine Tuning in Large Language Models. In *Proceedings of the 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 823–829. IEEE.

Hee, M. S.; Gao, Z.; Wang, Y.; Chu, X.; Lee, R. K.-W.; and Qin, Z. 2025. Contrastive Instruction Fine-Tuning Large Multimodal Model for Hateful Meme Classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 760–773.

Hee, M. S.; Kumaresan, A.; and Lee, R. K.-W. 2024. Bridging Modalities: Enhancing Cross-Modality Hate Speech Detection with Few-Shot In-Context Learning. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7785–7799. Miami, Florida, USA: Association for Computational Linguistics.

Hee, M. S.; and Lee, R. K.-W. 2025. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 774–785.

Hee, M. S.; Sharma, S.; Cao, R.; Nandi, P.; Nakov, P.; Chakraborty, T.; and Lee, R. K.-W. 2024. Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4407–4419. Miami, Florida, USA: Association for Computational Linguistics.

Hsieh, C.-Y.; Li, C.-W.; Yeh, C.-K.; Nakhost, H.; Fujii, Y.; Ratner, A.; and Pfister, T. 2023. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Hu, Y.; Hee, M. S.; Nakov, P.; and Lee, R. K.-W. 2025. Toxicity Red-Teaming: Benchmarking LLM Safety in Singapore's Low-Resource Languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 12194–12212.

Jaskie, K.; and Spanias, A. 2019. Positive and Unlabeled Learning Algorithms and Applications: A Survey. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–8. IEEE.

Karamanolakis, G.; Mukherjee, S.; Zheng, G.; and Awadallah, A. 2021. Self-training with weak supervision. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 845–863.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2611–2624.

Kiryo, R.; Niu, G.; du Plessis, M. C.; and Sugiyama, M. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, volume 30.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*.

Lu, J.; Ma, K.; Wang, K.; Xiao, K.; Lee, R. K.-W.; Xu, B.; Yang, L.; and Lin, H. 2025. Is LLM an Overconfident Judge? Unveiling the Capabilities of LLMs in Detecting Offensive Language with Annotation Disagreement. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 5609–5626. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Madukwe, K. J.; Gao, X.; and Xue, B. 2022. Token Replacement-Based Data Augmentation Methods for Hate Speech Detection. *World Wide Web*, 25(3): 1129–1150.

Mei, J.; Chen, J.; Lin, W.; Byrne, B.; and Tomalin, M. 2023. Improving hateful meme detection via retrieval-guided contrastive learning. *arXiv preprint arXiv:2311.08110*.

Nirmal, A.; Bhattacharjee, A.; Sheth, P.; and Liu, H. 2024. Towards Interpretable Hate Speech Detection Using Large Language Model-Extracted Rationales. *arXiv preprint arXiv:2403.12403*.

Plaza-Del-Arco, F. M.; Molina-González, M. D.; Ureña-López, L. A.; and Martín-Valdivia, M. T. 2021. A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis. *IEEE Access*, 9: 112478–112489.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rizos, G.; Hemker, K.; and Schuller, B. 2019. Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 991–1000.

Sakai, T.; du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2017. Semi-supervised classification based on classification from positive and unlabeled data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2998–3006. PMLR.

Shu, S.; Lin, Z.; Yan, Y.; and Li, L. 2020. Learning from multi-class positive and unlabeled data. In *2020 IEEE International Conference on Data Mining (ICDM)*, 1256–1261. IEEE.

Sprague, Z.; Yin, F.; Rodriguez, J. D.; Jiang, D.; Wadhwa, M.; Singhal, P.; Zhao, X.; Ye, X.; Mahowald, K.; and Durrett, G. 2024. To CoT or Not to CoT? Chain-of-Thought Helps Mainly on Math and Symbolic Reasoning. *arXiv preprint arXiv:2409.12183*.

Tür, G.; Tür, D.; and Schapire, R. E. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2): 171–186.

Wang, H.; Hee, M. S.; Awal, M. R.; Choo, K. T. W.; and Lee, R. K.-W. 2023a. Evaluating GPT-3 generated explanations for hateful content moderation. *arXiv preprint arXiv:2305.17680*.

Wang, H.; Tan, R. Y.; and Lee, R. K.-W. 2025. Cross-Modal Transfer from Memes to Videos: Addressing Data Scarcity in Hateful Video Detection. In *Proceedings of the ACM on Web Conference 2025*, 5255–5263.

Wang, H.; Wang, Z.; and Lee, R. K.-W. 2025. HateClipSeg: A Segment-Level Annotated Dataset for Fine-Grained Hate Video Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 13304–13310. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

Wang, H.; Yang, T. R.; Naseem, U.; and Lee, R. K.-W. 2024a. MultiHateClip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 7493–7502. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.;

Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; Schiele, B.; and Xie, X. 2023b. FreeMatch: Self-Adaptive Thresholding for Semi-Supervised Learning. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Xiao, Y.; Hu, Y.; Choo, K. T. W.; and Lee, R. K.-W. 2024. ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6012–6025. Miami, Florida, USA: Association for Computational Linguistics.

Xu, S.; Li, X.; Wu, S.; Zhang, W.; Tong, Y.; and Loy, C. C. 2023. Dst-det: Simple dynamic self-training for open-vocabulary object detection. *arXiv preprint arXiv:2310.01393*.

Yang, X.; Lv, F.; Liu, F.; and Lin, G. 2023. Self-training vision language BERTs with a unified conditional model. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8): 3560–3569.

Yu, H.; Zuo, W.; and Peng, T. 2005. A new PU learning algorithm for text classification. In *Mexican International Conference on Artificial Intelligence*, 824–832. Springer, Berlin, Heidelberg.

Zhao, S.; Schulter, S.; Zhao, L.; Zhang, Z.; Suh, Y.; Chandraker, M.; and Metaxas, D. N. 2024. Taming self-training for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13938–13947.

Zou, Y.; Yu, Z.; Kumar, B. V. K. V.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 289–305.