JOURNAL OF INFORMATION SCIENCE AND ENGINEERING XX, XXXX-XXXX (2024)
DOI: XX.XXX.XXXX.XX

MSBI: A Multi-Task Unified Framework to Integrate Transformers and Annotation Disagreements for Sexism Detection on Social Media

YI-ZENG FANG[1], LUNG-HAO LEE[2+], JUINN-DAR HUANG[1]

[1] *Institute of Electronics, National Yang Ming Chiao Tung University*
[2] *Institute of Artificial Intelligence Innovation, National Yang Ming Chiao Tung University*
E-mails: joycefang1213.ee11@nycu.edu.tw; lhlee@nycu.edu.tw; jdhuang@nycu.edu.tw

There is an increasing need for moderation tools that consider not only what is written, but also the context in which it appears and who it targets. To address this need, we propose the Multi-task Social Bias Identifier (MSBI), a transformer-based model designed to detect sexism in social media. In contrast to prior work that considers sexism-related subtasks separately and typically ignores information about the annotators, MSBI is trained jointly on three closely related objectives: sexism identification, source intention classification, and sexism type categorization. Training these three objectives together allows the model to reuse information across tasks and to pick up on subtle aspects of context that might be missed in a single-task setting. In addition, we feed annotators' demographic attributes into the training process so that the model can reflect differences in how people perceive the same content and move closer to providing user-specific assessments of social bias. Experiments on the EXIST-2024 dataset, which consists of 10,034 tweets with demographic annotations, show that MSBI outperforms state-of-the-art baselines on all three tasks, obtaining ICM-Soft scores of 1.0944, -0.2543, and -1.1762. These results show that demographically-informed multi-task learning is a promising direction for modeling complex social biases and for developing more flexible and fair content moderation tools in multilingual settings.

Keywords: Sexism Detection, Social Bias, Multi-task Learning, Text Classification, Bias Interpretability

1. INTRODUCTION

Social media platforms such as Twitter, Instagram, and Facebook have changed how people talk in public by allowing real-time communication with large audiences [1]. At the same time, their open nature also makes it easy for harmful content to spread, including harassment, biased messages, and discriminatory speech. As these platforms play a bigger role in shaping public opinion and everyday discussion, there is a clear need for content moderation systems that are both reliable and inclusive. In this work, we focus on sexism as one ongoing and difficult problem in online spaces. Sexist content can be very direct, but it can also be subtle or hidden in jokes and everyday language. In both cases, it causes gender stereotypes and helps create unfriendly or hostile environments for women and other targeted groups [2]. Detecting this type of content is particularly

[+]Corresponding author

hard on platforms like Twitter, where posts are short, often lack context, and may rely on hints or indirect wording. These conditions make simple keyword filters or shallow checks insufficient and call for methods that look beyond only the surface of the text.

Interpreting sexism in online content is not straightforward. Judgments are influenced by social norms, cultural background, and demographic characteristics of the readers, so the same post can be read in very different ways. What appears neutral or even humorous to one group may be viewed as offensive or discriminatory by another [3]. For automated systems that try to find and label sexist language, this variety is a major challenge, especially on social media where many languages and user groups are involved. Many existing methods rely on single-task classifiers and tend to focus on surface patterns. This makes it difficult for them to model the nuanced, context-dependent nature of bias and to reflect the viewpoints of different user groups. In this study, we take a step toward addressing these issues by examining how demographic information can be woven into model architectures in a principled way, so that the resulting interpretations of bias are more context-aware and equitable. We also look at whether treating several related interpretive tasks jointly, deciding whether a post is sexist, inferring the speaker's intention, and assigning a type of sexism, can lead to better performance than tackling each task in isolation.

This study is organized around three main research questions. First, how can demographic information be incorporated into the model so that it helps interpret biases across different user groups? Second, can we improve the detection of sexism on social media by jointly modeling multiple sexism-related tasks? Third, to what extent does hierarchical inference across these tasks help uncover nuanced or implicit sexist content that might otherwise be missed? To address these questions, we propose the **Multi-task Social Bias Identifier (MSBI)**, a unified transformer-based framework developed for the EXIST-2024 challenge. MSBI improves sexism detection in two key ways: (1) it jointly models three interrelated tasks, including sexism identification, intent classification, and fine-grained categorization, within a multi-task learning framework, and (2) it integrates annotator demographic attributes, including gender, age, and ethnicity, directly into the model input, allowing the system to simulate how different groups or users may perceive the same content.

To evaluate our approach, we conduct an experiment on the EXIST (sEXism Identification in Social neTworks) [4, 5] , which provides a standardized benchmark for advancing sexism detection in multilingual settings. The 2024 edition uses the Learning with Disagreements (LwD) framework [6], which asks models to take annotator diversity into account and to recognize that people from different demographic groups may not agree on what counts as sexism. MSBI achieves state-of-the-art performance with ICM-Soft scores of 1.0944 (Task 1), -0.2543 (Task 2), and -1.1762 (Task 3). These results demonstrate the framework's effectiveness in capturing subtle and diverse forms of sexist expression while improving interpretability and inclusivity in automated moderation.

2. Literature Review

Being able to detect anomalous or harmful speech on social media is an important part of maintaining reasonably safe online environments. Research on cyberbullying,

hate speech, and fake news has explored a wide range of machine learning and NLP methods [7], while highlighting recurring challenges such as ambiguous language, strong dependence on context, and the rapidly evolving nature of online communication. In the areas of cyberbullying and hate speech detection, research has gradually shifted from simple keyword-based methods to neural network models. For instance, transformer-based models can capture subtle linguistic patterns and contextual cues that earlier approaches often failed to detect [8, 9]. In fake news detection, systems commonly combine multiple sources of information, including textual content, indicators of source credibility, and patterns of information diffusion, often leveraging transformer encoders together with external knowledge sources. However, sexism presents additional challenges. Its expression is frequently indirect, deeply tied to cultural and social context, and therefore difficult to capture using single-task or purely text-focused models [10].

To support research on sexism detection, several benchmark datasets have been introduced, including early English-language resources such as AMI [11] and HatEval [12], as well as the multilingual EXIST benchmark. The EXIST dataset defines multiple subtasks, including sexism identification, source intention classification, and fine-grained categorization of sexist content [4, 5]. These challenges motivate the need for models that can jointly capture multiple aspects of sexist discourse. By learning these related tasks simultaneously, the model can develop shared representations that better capture the layered and context-dependent nature of sexist discourse. More broadly, advances in automated content moderation have been largely driven by transformer-based models such as BERT and RoBERTa [13, 14], which provide strong semantic representations and support cross-lingual applications [15]. However, moderation of harmful content remains strongly influenced by cultural subjectivity and demographic bias.

Recent studies have therefore explored the Learning with Disagreement (LwD) paradigm [16], which explicitly models disagreement among annotators rather than forcing a single consensus label. This is particularly important for subjective tasks such as sexism detection, where annotators may interpret the same content differently due to cultural background, personal experience, or social norms. Under the LwD framework, annotator responses are represented as soft-label distributions that quantify disagreement among annotators. These soft labels provide richer supervisory signals than traditional single-label annotations and have been shown to improve model robustness and generalization in subjective classification tasks [6, 17]. The EXIST shared tasks have adopted this paradigm in recent editions to better capture the inherent subjectivity involved in identifying sexist content. Building on this paradigm, several systems in recent EXIST shared tasks have explicitly modeled annotator disagreement using soft-label training or annotator-aware architectures, showing that incorporating annotator variability can improve sexism detection performance. [18, 19, 20] Demography-aware approaches attempt to address this issue by incorporating annotator attributes into the modeling process [21]. Recent work has further demonstrated that integrating annotator information directly into model training can improve the detection of subjective or socially sensitive content [22]. In this work, we extend this idea by feeding annotator demographic factors directly into the model architecture. Rather than correcting for bias only after training, MSBI learns how different social groups interpret and judge potentially sexist content, resulting in a subject-aware and contextually adaptive moderation framework that better reflects the diversity of real-world online environments.

3. Multi-task Social Bias Identifier

In this section, we describe the end-to-end pipeline of our sexism detection system and explain how the Multi-task Social Bias Identifier (MSBI) is used as its main component. As shown in Figure 1, the system follows a structured pipeline with three main stages: **pre-processing**, **multi-task detection with MSBI**, and **post-processing**. This integrated design helps maintain consistency across the data, supports robust feature extraction, and produces outputs that are easier to interpret and more suitable for practical deployment.

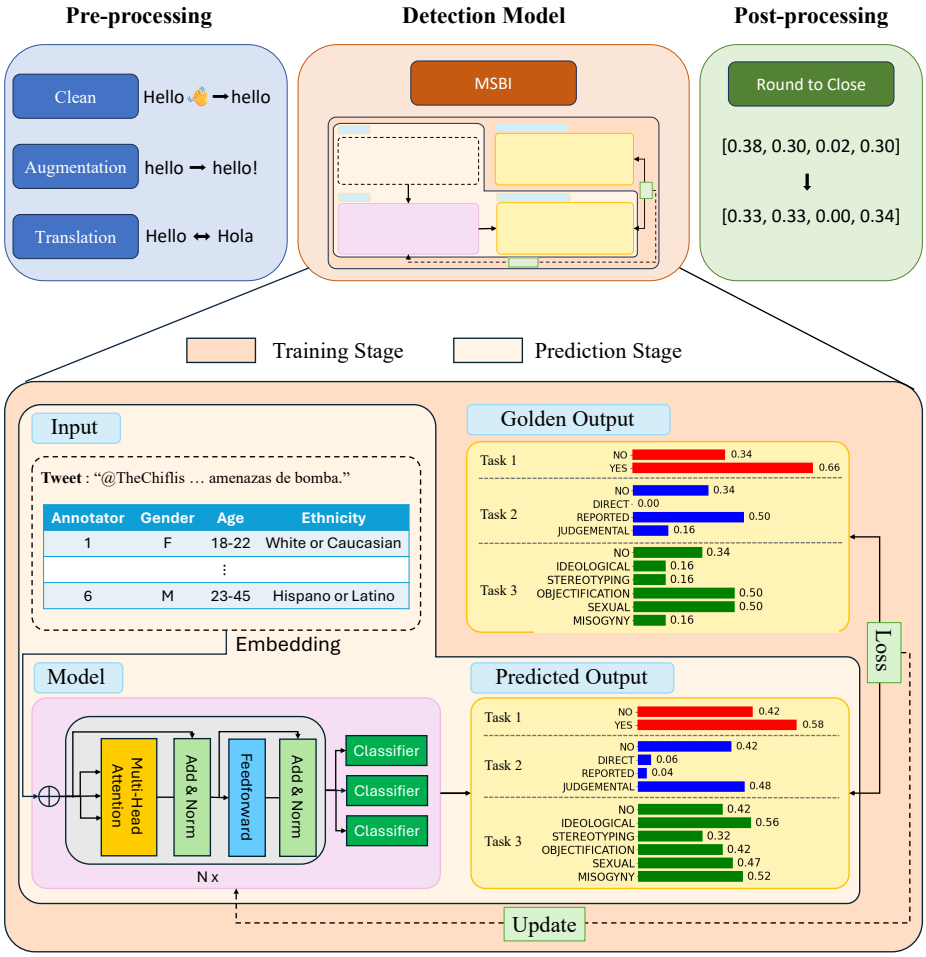


Fig. 1. Overview of the proposed system pipeline, including pre-processing, the MSBI-based detection model and post-processing. The lower panel illustrates the training and prediction stages, where tweet inputs and annotator information are encoded and used for multi-task sexism detection.

3.1 Pre-processing Pipeline

Data Cleaning A careful pre-processing step is crucial for obtaining reliable text input and reducing noise that does not help with sexism detection. In particular, we remove usernames, URLs, percentages, timestamps, dates, hashtags, and emojis, since these elements usually do not play a central role in annotators' judgments and rarely change the overall sentiment or intent of a tweet (see Figure 2). As a result, their semantic contribution cannot be reliably preserved and may instead introduce additional noise into the input representation. Stripping away such components also reduces spurious biases—for example, those linked to specific accounts or external links—and keeps the focus on the actual linguistic content of the message. In addition, one of the backbone models used in our framework is BERT [13], which adopts the WordPiece tokenizer. Tokens that cannot be properly segmented into known subword units, such as many emojis, are often mapped to the special token `[UNK]` (unknown). This mapping removes their original semantic information and may introduce unnecessary noise into the embedding space. To avoid this issue and maintain consistency in the input representation, we therefore remove emojis during pre-processing. We note, however, that emojis can carry meaningful semantic or affective cues [23, 24]. If a tokenizer or pretrained model is capable of preserving emoji semantics without converting them into unknown tokens, retaining emojis could potentially benefit sexism detection. Hashtags, emojis, and non-essential numerical tokens are often added for style or visibility rather than to convey meaning relevant to sexism classification. By filtering out these noninformative parts and removing emojis due to tokenizer limitations, we obtain a cleaner and more consistent dataset, allowing the model to concentrate on features that matter most for identifying sexist content and supporting more accurate downstream analyses.

Language	Original Sentence	Transform Sentence
Spanish	✊🏽 Desde Km0 nos solidarizamos con el llamado a la #MareaVerdePRen defensa de nuestros derechos sexuales y reproductivos. 📢 #AbortoLibreSeguroAccesible 🗓️ MAÑANA 28 de mayo ⏰ 2:00pm📌 Frente al Tribunal Federal del VSJ#AgitayTransforma#OtraSeguridadPublica https://t.co/fXxgJ6yeAD	desde km0 nos solidarizamos con el llamado a la defensa de nuestros derechos sexuales y reproductivos. mañana 28 de mayo 2:00pm frente al tribunal federal del vsj
English	Publication Alert 🚨 in @FeministPress WSQ https://t.co/rr0eWFJm60Caring is Love ❤️ Couple of years back, my work with @FRIDAfund caught me up with caring in young feminist movements with threats of backlash, burnout and violence. Speaking with @womyn_witch @Lady_karimah (1/2) https://t.co/HVNA9kFA0o	publication alert in wsq is love couple of years back, my work with caught me up with caring in young feminist movements with threats of backlash, burnout and violence. speaking with (1/2)

Fig. 2. Illustration of the data cleaning process, demonstrating the removal of usernames, URLs, hashtags, and other non-essential elements to focus on the core content.

AEDA-based Augmentation To increase the variety of training examples and improve robustness, we apply AEDA (An Easier Data Augmentation) [25]. In this approach, sentences are randomly split and punctuation marks from a small predefined set (".", ";",

Mode	Original Sentence	New Sentence
Spanish to English	la comunidad gamer es un antro de misóginos y supremacistas blancos hace años pero hoy día pocos se acuerdan de gamergate y todo lo que causó, y en cierto sentido sigue causando.	the gaming community has been a den of misogynists and white supremacists for years, but today few remember gamergate and everything it caused, and in a certain sense continues to cause.
English to Spanish	that's because they're bullys- plain and simple.don't understand it? don't want to understand it? don't want to think of a solution and actually fix the issue(s)?deflect and lie about actually taking action while finding a person or group you can further marginalize.	eso es porque son matones, simple y llanamente. ¿no lo entiendes? ¿no quieres entenderlo? ¿no quiere pensar en una solución y realmente solucionar el problema? fesvíe y mienta acerca de tomar medidas mientras encuentra a una persona o grupo al que puede marginar aún más.

Fig. 3. Bidirectional translation process, where English and Spanish texts are translated in both directions to introduce linguistic variability while preserving semantic integrity.

"?", ".", "!", ",") are inserted to form new variants of the original text, while preserving all original lexical items. AEDA is a widely used text data augmentation method in text classification research [26, 27]. One of its key advantages is its simplicity and semantic reliability, since it augments text without relying on additional models for rewriting. In contrast to more aggressive augmentation methods, such as synonym replacement, random insertion, swapping, or deletion [28], it does not remove or substitute existing tokens and therefore preserves the lexical cues that are critical for classification, although surface-level sentence segmentation may be affected by the inserted punctuation. This property is particularly important for sexism detection, where predictions often rely on specific lexical or semantic signals. Since AEDA retains these label-defining cues without modification, it is unlikely to distort the core information that annotators use to determine whether a tweet is sexist. As a result, the augmented samples remain close to the source data, and we obtain a richer training set that can help the model generalize better without adding substantial noise.

Bidirectional Translation We further adopt a translation-based augmentation strategy to enrich the dataset and support bilingual training. Specifically, we use the Google Translator API to translate each English tweet into Spanish and each Spanish tweet into English. This process doubles the dataset size and adds natural language variation (see Figure 3). The translated texts offer alternative phrasings of the same underlying content, allowing the model to observe how similar ideas are expressed in different languages and to pick up language-specific nuances. This bilingual expansion helps the model generalize more effectively across English and Spanish social media posts and improves its ability to detect and classify sexist content in multilingual settings.

3.2 MSBI Model Architecture

After pre-processing, the transformed text is passed to MSBI, the core analytical engine of the system. The MSBI architecture brings together the text of the tweet and the demographic information of the annotators, allowing the model to make predictions that are more sensitive to social context. The complete architecture is shown in Figure 1.

Incorporating Sentence and Demographic Representations Given an input tweet, we first tokenize the input and map each token to a d-dimensional embedding, resulting in $X_e \in \mathbb{R}^{l\times d}$ with l tokens. This embedded sequence is then processed by a pretrained transformer [29] model T, such as DeBERTaV3 [30], BERT [13], or RoBERTa [14], which outputs a representation $H = T(X_e) \in \mathbb{R}^{l\times d}$.

To transfer this representation into fixed-size sentence-level features, we apply both max-pooling and mean-pooling over the token dimension:

$$h_{\max} = \max(H) \in \mathbb{R}^d, h_{\text{mean}} = \text{mean}(H) \in \mathbb{R}^d. \tag{1}$$

The max pooling operation emphasizes the most informative features in the sequence, while the mean pooling operation preserves an overall summary of the sentence. To add sociocultural context, we include demographic information from the annotators, specifically gender, age, and ethnicity. These features allow the model to approximate how users with different social and cultural backgrounds might read the same content. Each attribute is first represented using one-hot encoding: let $g \in \{0,1\}^{C_g}$, $a \in \{0,1\}^{C_a}$, and $e \in \{0,1\}^{C_e}$ represent the one-hot vectors for gender, age, and ethnicity, respectively. These vectors are then projected into a common embedding space via trainable weight matrices:

$$v_g = W_g^\top g \in \mathbb{R}^{d_a}, v_a = W_a^\top a \in \mathbb{R}^{d_a}, v_e = W_e^\top e \in \mathbb{R}^{d_a}, \tag{2}$$

where $W_g \in \mathbb{R}^{C_g\times d_a}$, $W_a \in \mathbb{R}^{C_a\times d_a}$, and $W_e \in \mathbb{R}^{C_e\times d_a}$ are trainable matrices, and d_a is the dimension of the demographic embeddings.

Finally, we concatenate all information into a single feature vector:

$$z = \text{concat}(h_{\max}, h_{\text{mean}}, v_g, v_a, v_e) \in \mathbb{R}^{C_z}, \tag{3}$$

where $C_z = 2d + 3a$. This representation z encodes both the linguistic characteristics of the tweet and the interpretive context provided by the annotator demographics. This design choice enhances the system's sensitivity to subjective variation in social bias recognition, contributing to a more nuanced and inclusive content moderation framework.

Multi-Task Learning and Hierarchical Prediction MSBI is trained to solve three closely related classification tasks: (1) a binary task that predicts whether a tweet is sexist or not, (2) a multi-class task that identifies the underlying intent of the content, and (3) a multi-label task that assigns one or more specific sexism types. Although the label spaces differ, all three tasks are applied to the same tweet and are driven by overlapping social and linguistic cues. Importantly, the three tasks are not independent in the EXIST annotation scheme: Task 1 is a logical prerequisite for Task 2 and Task 3, and Task 3 further refines the semantic distinctions expressed in Task 2. This hierarchical dependency reflects a progression from coarse-grained detection to increasingly nuanced semantic interpretation.

Since these outputs stem from a shared interpretation of the tweet and its annotator context, we adopt a multi-task learning setup in which the model is optimized on all three tasks at once. This setup allows MSBI to learn a shared pool of features while still developing task-specific refinements for each objective, so that the tasks can support and

inform one another. Training them jointly helps the model generalize better by making use of the complementary signals that each task contributes. In particular, joint training exposes the model to correlated supervision signals that reflect the dataset's hierarchical label dependency, allowing it to implicitly internalize this structural bias through shared representations. This shared representation is especially important for subtle or implicit cases, where surface-level cues may be weak and fine-grained interpretation depends on a broader contextual understanding learned across tasks.

We implement *hard parameter sharing* [31], where a shared encoder processes the input and feeds into separate task-specific classification heads. Let $z \in \mathbb{R}^{2d+3d_a}$ be the joint representation of the tweet and demographic features, as defined in Section 3.2. Each classification head then applies a transformation to z:

$$o_1 = \text{softmax}(W_1^\top z + b_1), \qquad W_1 \in \mathbb{R}^{C_z \times t_1}, \quad t_1 = 2, \tag{4}$$

$$o_2 = \text{softmax}(W_2^\top z + b_2), \qquad W_2 \in \mathbb{R}^{C_z \times (t_2+1)}, \quad t_2 = 3, \tag{5}$$

$$o_3 = \text{sigmoid}(W_3^\top z + b_3), \qquad W_3 \in \mathbb{R}^{C_z \times (t_3+1)}, \quad t_3 = 5. \tag{6}$$

Here, o_1 estimates the probability that a tweet is sexist, o_2 chooses one intent category out of three, and o_3 produces a five-dimensional vector that encodes the presence or absence of each sexism type. To make sure all tweets can be used consistently during training, including those that Task 1 classifies as non-sexist, we expand the label spaces of Task 2 and Task 3 with an additional "no" category. This lets the model explicitly learn to predict a "non-sexist" outcome for the downstream tasks when appropriate, and allows us to train the full architecture in an end-to-end pipeline. This design ensures that the model learns both (i) task-specific decision rules and (ii) cross-task correlations that mirror the dataset hierarchy, which is essential for applying structured constraints at inference time. A detailed description of the label sets for each task is provided in Section 4.3.

After obtaining the outputs from all three tasks, they are compared against the true labels to compute the training loss. The overall loss function $\mathscr{L}$ for the multi-task learning framework is defined as the weighted sum of individual task losses:

$$\mathscr{L} = \lambda_1 \mathscr{L}_1 + \lambda_2 \mathscr{L}_2 + \lambda_3 \mathscr{L}_3, \tag{7}$$

where λ_1, λ_2, and λ_3 are scalar weights used to balance the contributions of each task. In our experiments, we set all weights to 1 for simplicity. A detailed analysis of different weight configurations analysis is provided in Section 4.6.

By optimizing all heads on a shared underlying representation, MSBI can take advantage of common learning signals while still respecting the specific requirements of each task. This design makes training more efficient, acts as a form of regularization, and reduces parameter redundancy relative to training three separate models. Overall, joint training serves as the representation-learning mechanism, while the hierarchical dependency is later made explicit during inference to structure the decision process from coarse to fine-grained levels.

Hierarchical Prediction Strategy After learning a shared representation through multi-task training, MSBI employs a *hierarchical prediction strategy* to organize inference across the three tasks. The first task—Sexism Identification—acts as the gatekeeper:

if a tweet is predicted as non-sexist, the system directly assigns the "non-sexist" label for Task 2 (intent classification) and Task 3 (category recognition), skipping further computation. This explicit gating reflects the intrinsic dataset structure: a non-sexist tweet cannot logically belong to any downstream intent or sexism category. More importantly, it separates coarse-grained detection from fine-grained interpretation, ensuring that subtle distinctions are only explored once the presence of sexism is sufficiently supported.

If Task 1 assigns a high probability to a tweet being sexist, the model then consults Tasks 2 and 3, using the output of Task 1 as a guiding signal. For Task 2, we replace the 'non-sexist" score with the probability from Task 1 and rescale the remaining class scores so that they still sum to 1. Since class selection only depends on which score is largest, this adjustment preserves the ranking of the intent categories. By conditioning intent prediction on the likelihood of sexism, the model can allocate more discriminative capacity to distinguishing nuanced intent types, rather than diffusing probability mass across mutually incompatible outcomes.

For Task 3, a label is predicted as present when its score exceeds 0.5. In this case, we scale each label probability in proportion to the sexist probability from Task 1, so that the confidence in downstream labels reflects how certain the model is that the tweet is sexist in the first place. This mechanism helps uncover nuanced or implicit sexist content by structuring the reasoning process: once a tweet passes the coarse sexism threshold, the model focuses on identifying specific semantic patterns at a finer granularity. In contrast, without hierarchical inference, the model may produce weak or inconsistent fine-grained signals for borderline cases, making subtle sexism harder to interpret reliably.

Overall, this hierarchical structure improves logical consistency, reduces downstream false positives, and aligns predictions with a step-by-step reasoning process. Conceptually, the hierarchical flow mirrors human annotation: annotators first determine whether sexism is present and then interpret its intent and form. By explicitly encoding this progression, the model better captures nuanced distinctions while maintaining coherent label combinations. Because hierarchical inference can amplify upstream mistakes, we emphasize that Task 1 achieves strong performance in our experiments, which mitigates excessive error propagation.

3.3 Post-processing with Round to Closed Value

To make MSBI's predicted probabilities align with how humans actually label the data, we apply the Round to Closed Value (RCV) method [17]. The idea is to adjust the model's outputs so they better reflect the discrete patterns found in annotator responses, making the predictions easier to understand and closer to real labeling behavior. For Task 1 and Task 2, which are both mono-label tasks, the class probabilities must sum to 1. Given the number of annotators, we enumerate all valid empirical distributions they could produce (for instance, with six annotators in Task 1, a distribution such as $[1/6, 5/6]$ is possible). We then compute the cosine similarity between the model's predicted probability vector and each of these valid distributions, and select the combination with the highest similarity as the adjusted output. Task 3 is multi-label, so the probabilities across labels are not constrained to sum to 1. In this case, we adjust the method as follows: for each label, we first round its predicted probability to the nearest valid value based on the annotator counts. If the sum of these adjusted values is still below 1, we raise the probability of the next closest label until the total reaches or slightly goes over 1. In this way,

we maintain coherence in the multi-label setting while still producing probability vectors that remain realistic, interpretable, and aligned with the observed annotation patterns.

4. Experiments and Performance Evaluation

MSBI performance across multiple tasks related to sexism detection was evaluated through a comprehensive experimental setup. This section describes the evaluation metrics, experimental settings, and the detailed results obtained from the evaluation process. The experiments were designed to measure the efficacy of the MSBI framework in terms of its accuracy, generalizability, and computational efficiency, using a combination of state-of-the-art transformer models and multi-task learning techniques. We also provide a detailed ranking of the system's performance compared to other submissions to relevant benchmarking challenges.

4.1 Evaluation Metrics

We use two distinct evaluation paradigms: soft-soft evaluation and hard-hard evaluation, both of which are used for the EXIST-2024 challenge to address different aspects of model predictions and their alignment with ground truth annotations. These two metrics can also reflect the nuanced demands of sexism detection tasks, particularly in capturing both objective and subjective facets of content classification.

Information Content Measure (ICM) Our evaluation metrics are based on the Information Content Measure (ICM), an extension of Pointwise Mutual Information (PMI) [32] tailored to multi-label and hierarchical classification scenarios. ICM assesses similarity by measuring the shared information between system predictions and ground truth labels, making it especially well-suited for complex classification tasks with overlapping or hierarchical categories. The general form of ICM is:

$$ICM(A,B) = \alpha_1 IC(A) + \alpha_2 IC(B) - \beta IC(A \cup B), \tag{8}$$

where $IC(A)$ represents the Information Content of the set A, and the parameters α_1, α_2, and β control the weighting of each term. When these parameters are set to 1, ICM behaves like PMI, offering flexibility to suit various classification needs. This flexibility makes ICM a valuable tool for hierarchical, multi-label classification tasks, where it has shown advantages over traditional evaluation metrics [33]. For hierarchical classification, ICM can be adapted as follows:

$$ICM(s(d), g(d)) = 2IC(s(d)) + 2IC(g(d)) - 3IC(s(d) \cup g(d)), \tag{9}$$

where $s(d)$ represents the system-predicted categories for document d and $g(d)$ denotes the ground truth categories. When predictions and ground truth labels align exactly, the ICM score is equal to the Information Content of $g(d)$; if no categories are assigned, the score defaults to $-IC(g(d))$, which serves as a normalization boundary. This metric setup helps to evaluate the model's capacity to match or approximate the true category structure, particularly valuable in hierarchical contexts.

Hard-Hard Evaluation To establish a traditional baseline for evaluation, the hard-hard approach assigns a single label to each instance based on the majority vote of annotators. This method assumes that each instance has a definitive, correct label, allowing for a straightforward comparison of the model's predictions to this fixed standard. Ground truth labels are derived using task-specific thresholds: in Task 1, a label is assigned if more than three annotators agree; in Task 2, if more than two annotators agree; and in multi-label Task 3, if more than one annotator agrees. Instances without a clear majority are excluded from the analysis if no label meets the threshold. Using the original ICM metric, this evaluation quantifies the similarity between predicted and ground truth labels, providing a rigorous measure of the model's accuracy under conventional classification assumptions.

Soft-Soft Evaluation To account for the variability in human interpretation, particularly in subjective tasks like sexism detection, we adopt a soft-soft evaluation approach with ICM-Soft, which is an adaptation of the Information Content Measure (ICM) that compares soft model outputs to soft ground truth labels. ICM-Soft measures how closely the model's probability distributions align with those derived from the distribution of annotator responses, reflecting a spectrum of human opinions rather than a single correct answer. This approach aligns with the LwD framework [6], which acknowledges that nuanced tasks may result in multiple valid interpretations.

In this framework, the Information Content (IC) of assigning category c with an agreement level v is defined as:

$$IC(\{\langle c, v\rangle\}) = -\log_2(P(\{d \in D : g_c(d) \geq v\})), \tag{10}$$

where $P(\{d \in D : g_c(d) \geq v\})$ represents the probability of instances in the gold standard meeting or exceeding the agreement level v for category c. This IC calculation considers the mean and deviation of agreement levels across instances for each class, applying the cumulative probability over an inferred normal distribution. This method ensures that the model's outputs are evaluated not only for accuracy but also in terms of how well they reflect the diversity seen in annotator responses, which is particularly valuable in culturally and context-sensitive tasks.

4.2 Settings

The models were trained with the *AdamW* optimizer (learning rate 1e-5) and a dropout rate of 0.1 to reduce overfitting. Training ran for 30 epochs with a maximum sequence length of 128 tokens. We used a batch size of 20 to balance gradient stability and hardware limits. All experiments were conducted on a single *NVIDIA Tesla V100 GPU*. Experimental results are based on scores reported in the official EXIST-2023 [4] and EXIST-2024 [34] leaderboards. Our model was trained exclusively on soft labels, with hard labels generated post-training. Soft labels reflect the distribution of annotator opinions, while the derived hard labels offer simplified majority-aligned classifications.

4.3 Datasets

Our experiments use the EXIST-2024 dataset [5, 34], which includes 10,034 tweets in English and Spanish annotated for sexism detection. The corpus is divided into training,

Table 1. Soft and hard labels for each task in the dataset. Soft label values are calculated as the average across the annotations provided by 6 annotators, representing the distribution of responses. Hard labels are determined by selecting the label that appears most frequently among the 6 annotator responses.

Task 1	Labels	['YES', 'YES', 'NO', 'YES', 'YES', 'YES']
	Soft labels	['NO': 0.16, 'YES': 0.84]
	Hard labels	'YES'
Task 2	Labels	['REPORTED', 'JUDGEMENTAL', '-', 'REPORTED', 'JUDGEMENTAL', 'REPORTED']
	Soft labels	['NO': 0.16, 'DIRECT': 0.0, 'REPORTED': 0.5, 'JUDGEMENTAL': 0.34]
	Hard labels	'REPORTED'
Task 3	Labels	[['OBJECTIFICATION'], ['OBJECTIFICATION', 'SEXUAL-VIOLENCE'], ['-'], ['STEREOTYPING-DOMINANCE'], ['SEXUAL-VIOLENCE'], ['IDEOLOGICAL-INEQUALITY', 'MISOGYNY-NON-SEXUAL-VIOLENCE']]
	Soft labels	['NO': 0.16, 'IDEOLOGICAL-INEQUALITY': 0.16, 'STEREOTYPING-DOMINANCE': 0.16, 'OBJECTIFICATION': 0.34, 'SEXUAL-VIOLENCE': 0.34, 'MISOGYNY-NON-SEXUAL-VIOLENCE': 0.16]
	Hard labels	['OBJECTIFICATION', 'SEXUAL-VIOLENCE']

development, and test sets, and the number of examples in each language is kept roughly balanced. Each tweet is labeled with respect to three related tasks that, taken together, describe how sexism manifests on social media:

- **Task 1: Sexism Identification** involves binary classification, determining whether a tweet contains sexist content.
- **Task 2: Source Intention** categorizes the intention behind a sexist tweet as either *Direct* (explicitly promoting sexism), *Reported* (sharing a personal experience), or *Judgmental* (criticizing sexist behavior).
- **Task 3: Sexism Categorization** is a multi-label task that assigns one or more categories to sexist content, including *ideological inequality*, *stereotyping dominance*, *objectification*, *sexual violence*, and *misogyny-related non-sexual violence*.

Each tweet was independently annotated by six individuals whose demographic attributes, including gender, age, ethnicity, education, and nationality. These annotations are used to derive both *hard labels* (majority consensus) and *soft labels* (distribution of

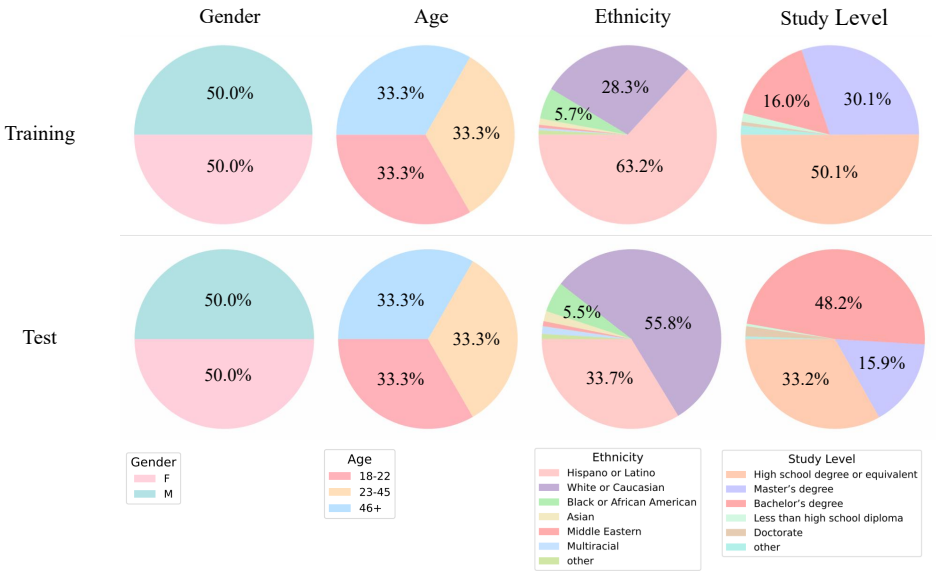


Fig. 4. Distribution of annotator demographics across training and test Sets.

annotator responses), as shown in Table 1. This annotation scheme allows the dataset to reflect the inherent subjectivity of sexism perception, supporting learning under the LwD paradigm.

To support the LwD setting adopted in EXIST-2024, the dataset maintains consistent annotator demographics across training and test splits (see Figure 4). Each split preserves a 50% gender balance and equal representation across three age groups (18–22, 23–45, 46+), each comprising roughly one-third of the total. This controlled distribution facilitates the study of subjective variation in sexism perception and enables systematic evaluation of demographically-aware modeling strategies. The stable improvements observed when incorporating demographic information further suggest that these features contribute complementary contextual signals rather than leading to demographic-specific overfitting. We emphasize that these demographic attributes are used in our framework to simulate diverse user viewpoints under the LwD paradigm, rather than as a mandatory requirement for real-world deployment.

This consistency helps the dataset capture a range of social perspectives and makes it easier to learn from annotator disagreement. Including viewpoints from different groups gives a strong foundation for training models that interpret sexist content with cultural and contextual awareness, supporting the development of more inclusive moderation systems.

4.4 Compared Methods

We used the following models for performance comparisons.

1) **ROH-NEIL** [35]

This approach uses multilingual transformer-based models for sexism detection

in English and Spanish tweets, focusing on both binary and multi-class classification tasks. The method uses cross-lingual transformers, particularly the XLM-RoBERTa-large model [36]. The model is trained using a combination of hard and soft-label classification, using Optuna for optimization [37]. This study also explores various loss functions, including binary cross-entropy and categorical cross-entropy, to optimize classification performance across hierarchical tasks.

2) **SINAI** [22]

This approach combines transformer-based text encoders with annotator socio-demographic information to improve sexism detection. It includes (1) a baseline model that fine-tunes multilingual transformers such as mDeBERTa and XLM-RoBERTa for text classification, and (2) a multimodal variant that integrates textual representations with annotator metadata through a Transformer-with-Tabular framework. By fusing linguistic features and demographic attributes, the method aims to model subjective variation in sexism perception.

3) **AI-UPV** [17]

This approach employs an ensemble strategy based on multilingual large language models to address sexism detection under the LwD framework. Specifically, it combines mBERT and XLM-RoBERTa to generate both soft labels, reflecting annotator disagreement, and corresponding hard labels. By aggregating predictions across models and applying distribution-level adjustments, the method aims to better align outputs with annotator consensus patterns.

4) **RMIT-IR** [38]

This approach leverages zero-shot in-context learning (ICL) [39] with large pre-trained language models for sexism characterization, and multimodal contrastive learning for meme-based sexism identification. Instead of task-specific fine-tuning, the method relies on structured prompting strategies to generate both soft and hard labels. It further incorporates annotator attributes, such as education level and gender, into prompt design to account for perspectivism in sexism interpretation.

5) **Awakened** [40]

This method adopts a Mixture of Transformers framework that combines multiple transformer models according to their language-specific strengths. It integrates general-purpose, task-specific, and data-source-specific encoders, selecting suitable models for English and multilingual inputs. In addition, the system applies a hierarchical prediction pipeline in which outputs from binary sexism detection guide downstream predictions for source intention and sexism categorization.

6) **ABCD** [19]

This approach uses Large Language Models (LLMs) in an ensemble architecture to tackle identify sexism in social network content. The method employs LLaMA 2 [41], mT5 [42], and XLM-RoBERTa, using fine-tuning and prompt engineering to adapt to classification tasks under the LwD paradigm. The ensemble consists of six independent models, each trained on distinct annotator-specific subsets, simulating human annotation variations.

Table 2. Performance comparison of Task 1 across models, evaluating ICM-Soft and ICM-Hard scores for all data, as well as Spanish (ES) and English (EN) subsets.

Method	All		ES		EN	
	ICM-Soft↑	ICM-Hard↑	ICM-Soft↑	ICM-Hard↑	ICM-Soft↑	ICM-Hard↑
Golden	3.1182	0.9948	3.1177	0.9999	3.1141	0.9798
ROH-NEIL	-2.8848	0.5795	-2.8671	0.5912	-2.9063	0.5521
SINAI	0.9030	0.5584	0.9527	0.5484	0.7984	**0.6004**
AI-UPV	0.7343	0.5119	0.7415	0.5069	0.6700	0.5211
RMIT-IR	-0.0011	0.4802	0.1840	0.4905	-0.2873	0.4502
Awakened	0.7196	0.5196	0.6563	0.5043	0.8171	0.5305
ABCD	0.9291	0.5957	0.9817	0.6166	0.8188	0.5561
GrootWatch	1.0538	0.5732	1.0498	0.5822	1.0106	0.5458
FHSTP	0.7852	0.5610	0.7824	0.5573	0.7484	0.5522
CLiC	0.7386	0.5059	0.7352	0.5294	0.6891	0.4591
MSBI (ours)	**1.0944**	**0.5973**	**1.1434**	**0.6215**	**1.0158**	0.5564

7) **GrootWatch** [43]

This approach introduces a context-enriched multi-task framework. For tweet classification, it employs a multi-headed XLM-RoBERTa architecture that integrates external contextual information, annotator demographic embeddings, and soft-label learning to model annotator disagreement. The model follows a hierarchical design in which a shared encoder supports multiple classification heads for sexism identification, intention detection, and multi-label categorization, with training optimized directly on soft-label distributions.

8) **FHSTP** [44]

This approach proposes Speech Concept Bottleneck Models (SCBM) for sexism detection, emphasizing interpretability. It uses Llama 3.1-8B-Instruct [45] to convert tweets into adjective-based concept representations that serve as bottleneck features for classification. A hybrid variant further combines these concept-level features with contextual embeddings from XLM-RoBERTa. The framework also considers demographic conditioning under the LwD paradigm.

9) **CLiC** [46]

This approach combines transformer fine-tuning and prompt optimization. For binary sexism identification, it fine-tunes language-specific transformers while incorporating annotator demographic information and training on soft-label distributions to model subjective perceptions. For intention detection and sexism categorization, it adopts prompt-based optimization with large language models, exploring annotator-aware conditioning and reasoning strategies. Predictions from Task 1 are propagated to downstream tasks in a hierarchical manner.

4.5 Results

For **Task 1 (Sexism Identification)**, MSBI achieved the best ICM-Soft score (1.0944), surpassing GrootWatch's 1.0538 with a 1.97% error reduction. Its ICM-Hard

Table 3. Performance comparison of Task 2 across models, evaluating ICM-Soft and ICM-Hard scores for all data, as well as Spanish (ES) and English (EN) subsets.

Method	All		ES		EN	
	ICM-Soft↑	ICM-Hard↑	ICM-Soft↑	ICM-Hard↑	ICM-Soft↑	ICM-Hard↑
Golden	6.2057	1.5378	6.2431	1.6007	6.1178	1.4449
ROH-NEIL	-5.7590	0.3883	-5.7499	**0.4783**	-5.7284	0.2664
SINAI	-2.2900	-0.0496	-2.0845	-0.1127	-2.7033	-0.0107
AI-UPV	-1.6836	0.1951	-1.6145	0.1827	-1.9067	0.1942
RMIT-IR	-4.5481	0.0855	-5.7527	0.1360	-6.1443	0.0152
Awakened	-2.0091	0.1812	-1.8428	0.1767	-2.4664	0.1629
ABCD	-1.8462	**0.4059**	-1.7958	0.4134	-1.8503	**0.3790**
GrootWatch	-0.5066	0.3266	-0.4294	0.4191	-0.6476	0.1929
FHSTP	-1.7928	0.0953	-1.5388	0.1329	-2.3042	0.0184
CLiC	-8.5179	-0.0969	-10.1213	-0.2902	-7.4727	0.0920
MSBI (ours)	**-0.2543**	0.3522	**-0.0756**	0.4457	**-0.5041**	0.2672

Table 4. Performance comparison of Task 3 across models, evaluating ICM-Soft and ICM-Hard scores for all data, as well as Spanish (ES) and English (EN) subsets.

Method	All		ES		EN	
	ICM-Soft↑	ICM-Hard↑	ICM-Soft↑	ICM-Hard↑	ICM-Soft↑	ICM-Hard↑
Golden	9.4686	2.1533	9.6071	2.2393	9.1255	2.0402
ROH-NEIL	-6.6622	**0.4433**	-5.6905	**0.4756**	-7.9493	**0.3974**
SINAI	-7.1306	0.1472	-7.0945	0.1491	-7.2312	0.1217
AI-UPV	-2.3183	-0.1862	-2.5779	-0.2284	-2.0487	-0.1528
RMIT-IR	-7.2098	-0.0344	-6.7226	0.0006	-7.8798	-0.0598
Awakened	-4.0748	-0.0042	-4.3009	-0.0365	-3.6903	0.0153
ABCD	-3.5160	0.3713	-3.0498	0.4495	-3.3702	0.2704
GrootWatch	-1.2566	0.2829	-1.1726	0.3395	-1.4405	0.2045
FHSTP	-11.2121	0.0978	-10.7122	0.1566	-11.9457	0.0140
CLiC	-9.6890	-0.1568	-8.6906	-0.1822	-11.0007	-0.1453
MSBI (ours)	**-1.1762**	0.3069	**-1.1280**	0.3138	**-1.2583**	0.2820

score (0.5973) was also slightly higher than ABCD's 0.5957, showing competitive performance even without explicit optimization for the hard metric (see Table 2). For **Task 2 (Source Intention)**, MSBI reached the top ICM-Soft score of -0.2543, an 3.76% error reduction over GrootWatch (-0.5066). Although not the highest in ICM-Hard (0.3522 vs. ABCD's 0.4059), MSBI still performs strongly in distinguishing intent categories (see Table 3). For **Task 3 (Sexism Categorization)**, MSBI obtained the best ICM-Soft score (-1.1762), a 0.75% error reduction over GrootWatch (-1.2566). Its ICM-Hard score (0.3069), while lower than ROH-NEIL's 0.4433, remains competitive given the complexity of multi-label sexism classification (see Table 4). In the tables, **Golden** denotes the upper-bound performance obtained under a fully correct prediction setting, i.e., the maximum achievable score when all instances are classified correctly.

Overall, MSBI's multi-task design effectively exploits the relationships among the three tasks, enabling superior performance compared to approaches that handle them

Table 5. Impact of incremental techniques on ICM-Soft scores across three tasks in the ablation study, showing performance improvements with the addition of data augmentation (DA), annotator demographic information (AI), Round to Closed Value (RC), and data translation (Tr.).

ICM-Soft All	Task 1 ↑	Task 2 ↑	Task 3 ↑
Baseline	0.7849	-1.2073	-3.2058
+ DA	0.8287	-0.9153	-2.9269
+ DA + AI	0.9410	-0.7256	-2.5230
+ DA + AI + RC	0.9862	-0.5597	-1.9450
+ DA + AI + RC + Tr.	**1.0084**	**-0.5208**	**-1.8042**

independently. By integrating annotator demographic information directly into the model—without additional architectures—MSBI captures diverse perceptions of sexism more faithfully. Its strong ICM-Soft improvements and stable ICM-Hard results demonstrate robust adaptability and state-of-the-art-level performance across ICM-Soft evaluation settings.

4.6 Discussion

Pipeline Design Table 5 summarizes how performance changes as we progressively extend the baseline model with additional components. Using five-fold cross-validation on the combined training and development sets, we observe steady gains on all three tasks when we add data augmentation (DA), annotator demographic information (AI), the Round to Closed Value method (RC), and data translation (Tr.). Taken together, these results highlight the benefit of a multi-faceted strategy for improving robustness and accuracy. Data augmentation is the first component that reliably improves the model, making it less sensitive to small changes in the input and boosting performance across all tasks. Adding annotator demographic information then allows the model to better mirror subjective judgments, which is particularly valuable for tasks that require fine-grained interpretation, such as source intention. The RC procedure further tightens the link between predicted probabilities and the actual distribution of annotator responses, yielding more noticeable gains in the multi-label sexism categorization task. On top of this, translation-based augmentation helps the model generalize in a bilingual setting, offering an additional improvement in overall performance. Taken together, these techniques each provide a modest but consistent benefit, and in combination they lead to more accurate and robust predictions across the three tasks.

Demographic Information of Annotators Table 6 further illustrates the impact of demographic information on model performance. Gender, age, and ethnicity each help sharpen the predictions across all three tasks, and using all three together yields the best ICM-Soft scores: 0.8287 for Task 1, -1.1771 for Task 2, and -3.0957 for Task 3. By contrast, adding education level, either on its own or alongside other demographic attributes, does not lead to additional gains. This is consistent with our decision to leave education out of the main model, since its uneven distribution across datasets may in-

Table 6. Impact of different combinations of annotator demographic information (gender, age, ethnicity, education level) on ICM-Soft scores across three tasks.

Gender	Age	Ethnicity	Education Level	Task 1 ↑	Task 2 ↑	Task 3 ↑
				0.7849	-1.2073	-3.2058
			✓	0.7748	-1.2108	-3.2095
		✓		0.7896	-1.1953	-3.0867
	✓			0.7835	-1.1934	-3.1224
✓				0.7856	-1.2057	-3.1503
		✓	✓	0.7892	-1.2147	-3.2083
	✓		✓	0.7892	-1.2077	-3.1820
✓			✓	0.7849	-1.2187	-3.2207
	✓	✓		0.8005	-1.1953	-3.1806
✓		✓		0.7903	-1.2031	-3.1820
✓	✓			0.7919	-1.2058	-3.1974
	✓	✓	✓	0.8039	-1.1934	-3.1804
✓		✓	✓	0.8060	-1.1978	-3.1502
✓	✓		✓	0.8057	-1.1953	-3.1469
✓	✓	✓		**0.8287**	**-1.1771**	**-3.0957**
✓	✓	✓	✓	0.8052	-1.1878	-3.1202

Table 7. Sensitivity analysis of different loss weight ratios on ICM-Soft scores across three tasks.

Weight Ratio (T1:T2:T3)	Task 1 ↑	Task 2 ↑	Task 3 ↑
1:1:1	1.0084	-0.5208	-1.8042
2:1:1	**1.1167**	-0.6534	-1.9451
1:2:1	0.9654	**-0.4872**	-1.9672
1:1:2	0.9455	-0.6375	**-1.7487**

troduce unwanted variation and affect stability. Focusing on gender, age, and ethnicity therefore provides a more stable demographic basis and reduces the risk of biases arising from inconsistent educational backgrounds.

Loss Weight Sensitivity Analysis To examine the impact of loss weighting, we evaluated four representative configurations: equal weighting (1:1:1) and three asymmetric settings emphasizing one task at a time (2:1:1, 1:2:1, and 1:1:2). Table 7 shows that increasing the weight of a specific task consistently improves its own ICM-Soft score. For example, assigning a higher weight to Task 1 (2:1:1) raises its score from 1.0084 to 1.1167. Similar trends are observed when Task 2 or Task 3 is emphasized. However, this improvement comes with a clear trade-off: enhancing one task systematically degrades

Table 8. Comparison between MSBI and recent LLM-based approaches on ICM-Soft scores across three tasks.

Method	Task 1 ↑	Task 2 ↑	Task 3 ↑
LLRP	0.8256	-0.7371	-1.9946
EGDP	0.7562	-0.8269	-2.2705
MSBI (ours)	**1.0084**	**-0.5208**	**-1.8042**

the performance of the others. The asymmetric settings lead to noticeable declines in at least one remaining task, indicating that over-prioritizing a single objective may reduce the effectiveness of shared representations in the multi-task framework. In contrast, the equal-weight configuration achieves the most balanced performance across all three tasks. Although it does not produce the highest score for any single task, it avoids substantial degradation and yields the most stable overall behavior. Therefore, we retain the equal-weight setting as it better supports the unified modeling objective of MSBI.

Comparison with Recent LLM-based Approaches Recent studies have explored LLM-centered strategies for sexism detection from different perspectives. The first line of work, which we refer to as *LLM Low-Resource Prompting* (LLRP), systematically evaluates zero-shot, few-shot, and reasoning-based prompting, as well as lightweight fine-tuning and data augmentation, for sexism detection in a low-resource language setting [47]. Their findings highlight the effectiveness of fine-tuned LLMs for binary detection, while also noting challenges in more fine-grained categorization. The second line, which we refer to as *Expert-Guided Definition Prompting* (EGDP), investigates how expert–LLM interaction and collaboratively refined definitions can enhance zero-shot sexism detection through improved prompt design [48]. This approach emphasizes definitional clarity and interaction strategies rather than architectural modifications. In contrast to both LLRP and EGDP, MSBI is not built around prompt engineering or zero-shot inference. Instead, it adopts a demographically informed multi-task architecture with hierarchical inference, jointly modeling sexism identification, source intention, and fine-grained categorization within a unified framework. By explicitly incorporating annotator demographic attributes and enforcing cross-task consistency during inference, MSBI captures subjective variation and inter-task dependency more directly. As shown in Table 8, this design leads to stronger performance across Task 1–Task 3, suggesting that integrating demographic context and structured multi-task learning provides more stable and comprehensive improvements than relying solely on prompt-based LLM strategies.

5. Conclusion and Future Work

This study introduced the MSBI, a unified transformer-based system for advancing sexism detection in multilingual social media. MSBI jointly models three related tasks—**Sexism Identification**, **Source Intention Classification**, and **Sexism Categorization**—within a multi-task framework, enabling it to capture cross-task dependencies and improve understanding of both implicit and explicit sexist expressions. A key contri-

bution is the direct integration of annotator demographic attributes (gender, age, ethnicity) into the model. Instead of using demographics as post-hoc metadata, MSBI incorporates them during training to simulate how diverse social groups interpret potentially biased content. This aligns with the LwD paradigm in EXIST-2024 and allows the model to better reflect subjective, culturally dependent judgments.

Our experiments on EXIST-2024 show that MSBI reaches state-of-the-art ICM-Soft scores on all three tasks. Taken together, these gains suggest that combining multi-task learning with demographically-informed modeling offers clear benefits for building content moderation systems that are both more accurate and more inclusive.

Future work includes extending MSBI to additional platforms and modalities, expanding demographic modeling with user-centric signals, and exploring continuous learning to adapt to evolving online discourse. These directions aim to further strengthen MSBI's applicability in real-world moderation systems.

ACKNOWLEDGMENT

In preparing this paper, we used ChatGPT solely to aid in polishing the writing and improving clarity. The models were employed for grammar checking, smoothing sentence flow, and rephrasing for readability. All research ideas, methodological designs, experiments, and analyses were conceived and conducted entirely by the authors without reliance on LLMs. This work was partially supported by the National Science and Technology Council, Taiwan, under the grants NSTC 114-2221-E-A49-059-MY3 and NSTC 111-2628-E-A49-029-MY3.

REFERENCES

1. A. Keramatfar, H. Amirkhani, and A. Jalaly Bidgoly, "Multi-thread hierarchical deep model for context-aware sentiment analysis," *Journal of Information Science*, Vol. 49, no. 1, 2023, pp. 133–144.
2. P. Y. W. Myint, S. L. Lo, and Y. Zhang, "Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction," *Information Processing and Management*, Vol. 61, no. 4, 2024, pp. 1–20.
3. M. Z. U. Rehman, S. Zahoor, A. Manzoor, M. Maqbool, and N. Kumar, "A context-aware attention and graph neural network-based multimodal framework for misogyny detection," *Information Processing and Management*, Vol. 62, no. 1, 2025, pp. 1–20.
4. L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso, "Overview of EXIST 2023: sexism identification in social networks," in *Proceedings of the 2023 European Conference on Information Retrieval*, 2023, pp. 593–599.
5. L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, and D. Spina, "Overview of EXIST 2024 – learning with disagreement for sexism identification and characterization in social networks and

memes," in *Proceedings of the 2024 Conference and Labs of the Evaluation Forum*, 2024.

6. A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, "Learning from disagreement: A survey," *Journal of Artificial Intelligence Research*, Vol. 72, 2021, pp. 1385–1470.
7. L. Xing, S. Li, Q. Zhang, H. Wu, H. Ma, and X. Zhang, "A survey on social network's anomalous behavior detection," *Complex and Intelligent Systems*, 2024, pp. 1–16.
8. T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," *Information Processing and Management*, Vol. 60, no. 5, 2023, pp. 1–44.
9. A. Ghenai, Z. Noorian, H. Moradisani, P. Abadeh, C. Erentzen, and F. Zarrinkalam, "Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users," *Information Processing and Management*, Vol. 62, no. 3, 2025, pp. 1–28.
10. A. Vetagiri, P. Pakray, and A. Das, "A deep dive into automated sexism detection using fine-tuned deep learning and large language models," *Engineering Applications of Artificial Intelligence*, Vol. 145, 2025, pp. 1–17.
11. E. Fersini, D. Nozza, P. Rosso *et al.*, "Overview of the evalita 2018 task on automatic misogyny identification (ami)," in *Proceedings of the Working Notes of 2018 Conference and Labs of the Evaluation Forum*, Vol. 2263, 2018, pp. 1–9.
12. V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proceedings of the 2019 international workshop on semantic evaluation*, 2019, pp. 54–63.
13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, no. 2, 2019, pp. 4171–4186.
14. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
15. S. K. Pandey, H. S. Shekhawat, and S. Prasanna, "Multi-cultural speech emotion recognition using language and speaker cues," *Biomedical Signal Processing and Control*, Vol. 83, 2023, pp. 1–9.
16. A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, and M. Poesio, "Semeval-2021 task 12: Learning with disagreements," in *Proceedings of the 2021 international workshop on semantic evaluation*, 2021, pp. 338–347.
17. A. F. Magnossão de Paula, G. Rizzi, E. Fersini, and D. Spina, "AI-UPV at EXIST 2023–sexism characterization using large language models under the learning with disagreements regime," *arXiv preprint arXiv:2307.03385*, 2023.
18. Y.-Z. Fang, L.-H. Lee, and J.-D. Huang, "NYCU-NLP at exist 2024: Leveraging transformers with diverse annotations for sexism identification in social networks." in *Proceedings of the Working Notes of 2024 Conference and Labs of the Evaluation Forum*, 2024, pp. 1003–1011.

19. L. M. Quan and D. V. Thin, "Sexism identification in social networks with generation-based language models," in *Proceedings of the Working Notes of 2024 Conference and Labs of the Evaluation Forum*, 2024, pp. 1165–1176.
20. J. C. Prajogo, L.-H. Lee, and H.-I. Lin, "Nycu-nlp at exist 2025: An empirical study of annotator-aware two-stage pipeline for sexism detection in tweets," in *Proceedings of the Working Notes of 2025 Conference and Labs of the Evaluation Forum*, Vol. 4038, 2025, pp. 2119–2132.
21. J. Yang, A. A. Soltan, D. W. Eyre, Y. Yang, and D. A. Clifton, "An adversarial training framework for mitigating algorithmic biases in clinical machine learning," *NPJ Digital Medicine*, Vol. 6, no. 1, 2023, pp. 1–10.
22. M. E. Vallecillo-Rodríguez, F. M. Plaza-del Arco, L. A. Ureña-López, M. T. Martín-Valdivia, and A. Montejo-Ráez, "Integrating annotator information in transformer fine-tuning for sexism detection," in *Proceedings of the Working Notes of 2023 Conference and Labs of the Evaluation Forum*, 2023, pp. 1091–1106.
23. P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of emojis," *PloS one*, Vol. 10, no. 12, 2015, pp. 1–22.
24. R. Godard and S. Holtzman, "The multidimensional lexicon of emojis: A new tool to assess the emotional content of emojis," *Frontiers in psychology*, Vol. 13, 2022, pp. 1–10.
25. A. Karimi, L. Rossi, and A. Prati, "AEDA: an easier data augmentation technique for text classification," *arXiv preprint arXiv:2108.13230*, 2021.
26. Z. Liu, Z. Tang, X. Shi, A. Zhang, M. Li, A. Shrivastava, and A. G. Wilson, "Learning multimodal data augmentation in feature space," *arXiv preprint arXiv:2212.14453*, 2022.
27. S. Shi, K. Hu, J. Xie, Y. Guo, and H. Wu, "Robust scientific text classification using prompt tuning based on data augmentation with l2 regularization," *Information Processing and Management*, Vol. 61, no. 1, 2024, pp. 1–19.
28. J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.
29. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6000–6010.
30. P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," *arXiv preprint arXiv:2111.09543*, 2021.
31. S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.
32. K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, Vol. 16, no. 1, 1990, pp. 22–29.
33. E. Amigo and A. Delgado, "Evaluating extreme hierarchical multi-label classification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 5809–5819.
34. L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, and D. Spina, "Overview of EXIST 2024–learning with disagreement for sexism identification and characterization in tweets and memes (ex-

tended overview)," in *Proceedings of the 2024 Conference and Labs of the Evaluation Forum*, 2024.
35. R. Kooniredddy and N. Adel, "ROH_NEIL@ EXIST2023: Detecting sexism in tweets using multilingual language models." in *Proceedings of the Working Notes of 2023 Conference and Labs of the Evaluation Forum*, 2023, pp. 967–984.
36. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.
37. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "OPTUNA: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, pp. 2623–2631.
38. T. K. Smith, H. R. Nie, J. R. Trippas, and D. Spina, "RMIT-IR at EXIST lab at CLEF 2024," in *Proceedings of the Working Notes of 2024 Conference and Labs of the Evaluation Forum*, 2024, pp. 1237–1252.
39. Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, and Z. Sui, "A survey on in-context learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 1107–1128.
40. A. Petrescu, C.-O. Truică, and E.-S. Apostol, "Language-based mixture of transformers for EXIST2024," in *Proceedings of the Working Notes of 2024 Conference and Labs of the Evaluation Forum*, 2024, pp. 1157–1164.
41. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
42. L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.
43. N. Nowakowski, L. Calogiuri, E. Egyed-Zsigmond, D. Nurbakova, J. Erbani, and S. Calabretto, "Grootwatch at exist 2025: Automatic sexism detection on social networks-classification of tweets and memes," *Lecture Notes in Computer Science*, 2025.
44. R. Labadie-Tamayo, A. J. Böck, D. Slijepčević, X. Chen, A. Babic, and M. Zeppelzauer, "FHSTP@ exist 2025 benchmark: sexism detection with transparent speech concept bottleneck models," *arXiv preprint arXiv:2507.20924*, 2025.
45. A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
46. P. Pastells, M. Vázquez, M. Farrús, and M. Taulé, "CLiC at exist 2025: combining fine-tuning and prompting with learning with disagreement for sexism detection," *Proceedings of the Working Notes of 2025 Conference and Labs of the Evaluation Forum*, 2025.
47. L. S. M. Altın and H. Saggion, "Investigating large language models'(llms) capabilities for sexism detection on a low-resource language," in *Proceedings of the 2025 International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, 2025, pp. 771–779.

48. M. Reuver, I. Sen, M. Melis, and G. Lapesa, "Tell me what you know about sexism: Expert-llm interaction strategies and co-created definitions for zero-shot sexism detection," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 8438–8467.



Yi-Zeng Fang received the B.S. degree in Electrical Engineering from National Yang Ming Chiao Tung University, Taiwan. She is currently pursuing the Ph.D. degree in the Institute of Electronics at National Yang Ming Chiao Tung University, Taiwan. Her research interests include large language models.



Lung-Hao Lee received the Ph.D. degree in computer science and information engineering from National Taiwan University, in 2015. From 2015 to 2018, he was a Postdoctoral Fellow with National Taiwan Normal University. He was an assistant professor in the Department of Electrical Engineering at National Central University. He is currently an associate professor in the Institute of Artificial Intelligence Innovation at National Yang Ming Chiao Tung University. His research interests include natural language processing, web information retrieval, and artificial intelligence technologies.



Juinn-Dar Huang Juinn-Dar Huang received the B.S. and Ph.D. degrees in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1992 and 1998, respectively. He is currently a distinguished professor in the Department of Electronics Engineering and the Institute of Electronics, National Yang Ming Chiao Tung University. He once served as Deputy General Manager at the Program Office of AI Research, Taiwan, from 2017 to 2018, and as VP Research at Neuchips Corp., a Taiwan-based AI startup company, from 2021 to 2023. He received the Best Paper Award in the IEEE Computer Society Annual Symposium on VLSI (ISVLSI) 2011 and the IEEE International Symposium on VLSI Design, Automation and Test (VLSI-DAT) 2021. He was the Technical Program Chair of IEEE VLSI-DAT 2019, and the General Chair of IEEE VLSI-DAT 2020. His current research interests include accelerator design for machine learning and edge computing, the development and refinement of deep learning algorithms, and electronic design automation (EDA). Dr. Huang is currently a senior member of the IEEE and a member of the ACM, as well as a member of Phi Tau Phi.