IJWR

Building Safer Social Spaces: Addressing Body Shaming with LLMs and Explainable AI

Sajedeh Talebi, Neda Abdolvand*
Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran;
sajedeh.talebi1998@gmail.com , n.abdolvand@alzahra.ac.ir

ABSTRACT

This study tackles body shaming on Reddit using a novel dataset of 8,067 comments from June to November 2024, encompassing external and self-directed harmful discourse. We assess traditional Machine Learning (ML), Deep Learning (DL), and transformer-based Large Language Models (LLMs) for detection, employing accuracy, F1-score, and Area Under the Curve (AUC). Fine-tuned Psycho-Robustly Optimized BERT Pretraining Approach (Psycho-RoBERTa), pre-trained on psychological texts, excels (accuracy: 0.98, F1-score: 0.994, AUC: 0.990), surpassing models like Extreme Gradient Boosting (XG-Boost) (accuracy: 0.972) and Convolutional Neural Network (CNN) (accuracy: 0.979) due to its contextual sensitivity. Local Interpretable Model-agnostic Explanations (LIME) enhance transparency by identifying influential terms like "fat" and "ugly." A term co-occurrence network graph uncovers semantic links, such as "shame" and "depression," revealing discourse patterns. Targeting Reddit's anonymity-driven subreddits, the dataset fills a platform-specific gap. Integrating LLMs, LIME, and graph analysis, we develop scalable tools for real-time moderation to foster inclusive online spaces. Limitations include Reddit-specific data and potential misses of implicit shaming. Future research should explore multi-platform datasets and few-shot learning. These findings advance Natural Language Processing (NLP) for cyberbullying detection, promoting safer social media environments.

Keywords*— Body Shaming, Reddit, Machine Learning, Deep Learning, Large Language Models, Local Interpretable Model-agnostic Explanations, Content Moderation*

1. Introduction

Social media platforms like Reddit and X have transformed global communication by enabling fast information sharing and connectivity [1]. However, these platforms also promote harmful behaviors, such as body shaming, which is a form of bullying targeting physical features like weight, height, or skin conditions [2]. Body shaming leads to significant psychological harm, including anxiety, depression, and low self-esteem, especially among adolescents and those with body image issues, worsened by unrealistic beauty standards on social media [2-3]. Due to the vast amount of user-generated content, manually detecting body shaming is impractical, making the use of automated computational methods necessary. Though Machine Learning (ML) has improved the detection of online harassment, hate speech, and cyberbullying [4-5], body shaming remains underexplored as a separate form of harassment. To address this, this study has two main objectives: creating a dataset of 8,067 Reddit comments that reflect body shaming discourse and evaluating the effectiveness of traditional ML, Deep Learning (DL), and transformer-based Large Language Models (LLMs) for identifying this content, with a focus on accuracy and interpretability. We incorporate Local Interpretable Model-agnostic Explanations (LIME) to provide transparent model decisions and utilize a co-occurrence network graph to identify semantic patterns. The research aims to answer two questions: (1) How does a dedicated Reddit dataset improve body shaming detection models? and (2) How do transformer-based LLMs compare to traditional methods, and can LIME enhance moderation transparency? We evaluate fine-tuned and partially tuned LLMs alongside ML and DL models using standard metrics and an expert-labeled dataset. This study contributes by: (1) providing a new 8,067-comment Reddit dataset for body shaming, (2) assessing ML, DL, and LLMs for detection, (3) enhancing transparency with LIME, and (4) employing graph analysis to uncover semantic patterns.

http://dx.doi.org/10.22133/ijwr.2025.525312.1286
Citation S. Talebi, N. Abdolvand, " Building Safer Social Spaces: Addressing Body Shaming with LLMs and Explainable AI", *International Journal of Web Research*, vol.8, no.3,pp.59-72, 2025, doi: 10.22133/ijwr.2025.525312.1286.

*Coressponding Author
Article History: Received: 21 April 2025 ; Revised: 3 June 2025 ; Accepted: 17 June 2025.

By addressing both externally directed and self-directed body shaming, this research captures the entire range of harmful discourse and supports the development of safer, more inclusive online communities. This study builds upon our earlier work [6], which investigated the use of LLMs for identifying body shaming. It serves as a pioneering effort in utilizing Natural Language Processing (NLP) for analyzing social media content.

The structure of this paper follows: Section 2 reviews related work on computational approaches to body shaming and online harassment; Section 3 details the methodology, including data collection, preprocessing, and model selection; Section 4 presents the experimental results and LIME-derived insights; Section 5 discusses the findings and limitations; Section 6 outlines the contributions and practical implications; and Section 7 concludes with a summary and directions for future research.

2. Related Works

Body shaming on social media significantly harms mental health, contributing to issues like anxiety and low self-esteem. Studies highlight that users frequently discuss sensitive topics such as weight, body shape, and facial features on platforms like Reddit, often in derogatory ways, amplifying harmful discourse [7-9]. Increasing efforts have been made to detect such content using ML and resampling techniques, which effectively identify body-shaming language, offering valuable tools for intervention. Graph-based analysis of Reddit posts identifies key individuals, behaviors, and communication patterns, providing insights into how body shaming manifests in different groups. Topic modeling reveals how harmful attitudes toward body image are commonly expressed and normalized in these communities. Support Vector Machine (SVM) have been applied to detect body shaming, though often not treated as a specific class. Classical sentiment analysis using Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction and Naive Bayes (NB) models has detected body shaming in YouTube video logs [10-14]. Advancements in DL have enhanced cyberbullying detection across platforms. LLMs like Generative Adversarial Network-Bidirectional Encoder Representations from Transformers (GAN-BERT), Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), and Extreme Language Network (XLNet) have surpassed traditional ML models (e.g., SVM, Logistic Regression) in identifying body shaming, particularly on X [15]. The trigram model performed well in identifying positive instances of body shaming without sacrificing accuracy compared to TF-IDF [16]. A study combining Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and Gated Recurrent Unit (GRU) architectures achieved robust results in detecting Arabic cyberbullying [17]. Transformer models excelled due to their contextual understanding, with BERT achieving the highest F1-score and Matthews Correlation Coefficient on larger datasets, while RoBERTa performed best on smaller ones. Among traditional models, Random Forests showed strong performance [18]. Transformer models like ALBERTO, Cross-Lingual Language Model-Robustly Optimized BERT Pretraining Approach (XLM-RoBERTa), and Universal Multilingual BERT (Um-BERT) achieved 94% accuracy in classifying body-shaming content on Instagram, highlighting their versatility [19].

Unlike prior studies focusing on X and Instagram, our research examines Reddit's subreddit-driven discourse. We employ fine-tuned and partially tuned LLMs, such as Psycho-Robustly Optimized BERT Pretraining Approach and XLNet, to detect body-shaming language, integrating LIME for transparent content moderation. These transformer-based models excel in capturing complex linguistic nuances, enhancing detection accuracy. LIME provides clear insights into model decisions, fostering trust and supporting effective moderation practices, establishing a robust framework for addressing body shaming on social media.

2.1. Research Gap

Body shaming, a specific type of cyberbullying, is underexplored despite progress in detecting online harassment. Most studies focus on platforms like YouTube and X, using small datasets (fewer than 2,000 samples) that limit model generalizability and fail to capture diverse body shaming patterns [9-11]. Reddit, with its anonymous, subreddit-specific discussions, is often ignored, as prior research targets general hate speech without addressing body shaming's subtle language, including implicit and self-directed comments [13]. The absence of Reddit-specific datasets hampers effective model development. Moreover, advanced LLMs are underused, with studies relying on basic fine-tuning and treating models as black boxes, often neglecting interpretability tools like LIME and semantic analysis using DTM-based methods. This study addresses these gaps with a new 8,067-comment Reddit dataset and a comparative analysis of machine learning, deep learning, and transformer-based LLMs. We integrate LIME for transparent predictions and DTM-based term co-occurrence analysis to reveal semantic patterns in body shaming discourse. Our methodology tackles challenges like small dataset sizes, Reddit's unique environment, limited LLM use, and lack of model transparency, using LIME for clear decision-making insights and DTM to map semantic connections, setting a new benchmark for body shaming detection in NLP.

3. Materials and Methods

This study employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework [20], a widely adopted, iterative methodology that structures data mining projects (starting from business understanding to deployment). Next, during the modelling stage, a diverse set of ML, DL, and LLM models, both fine-tuned and partially tuned, are employed to classify comments as body shaming or non-body shaming. Figure 1 illustrates the proposed framework for body shaming detection using the CRISP-DM methodology.

Grounded in the CRISP-DM framework, as shown in the Figure above, this study adopts a structured, step-by-step approach to body shaming detection on Reddit. The process begins with data collection, where raw user comments related to various forms of body shaming, such as fat shaming, tall shaming, and skinny shaming, are gathered from relevant Reddit threads. In the data preprocessing phase, several text-cleaning techniques are applied to prepare the data for analysis. These include lowercasing, removal of irrelevant characters, tokenization, and stemming, all of which help standardise and refine the text for subsequent modelling. Next, during the modelling stage, a diverse set of MLs, DLs, and LLMs, both fine-tuned and partially tuned, is employed to classify comments as body shaming or non-body shaming. To enhance interpretability, LIME is used to assess feature importance, identifying keywords and phrases that significantly influence model predictions. To further explore the structure of body shaming discourse, a term co-occurrence network is constructed and visualised using a graph-based method. This reveals semantic patterns and relationships between commonly used terms within the dataset. Finally, the performance of each model is evaluated using multiple metrics, including Accuracy, F1-Score, and Area Under the Curve (AUC) [21], to ensure a comprehensive understanding of their predictive capability and practical utility. This systematic process results in a robust and interpretable framework for detecting body shaming content on Reddit, contributing valuable insights to the field of cyberbullying detection.

3.1. Data Collection and Annotation

This study aims to develop an effective methodology for detecting body shaming content on social media, addressing a critical form of online bullying that contributes to mental health issues such as depression, anxiety, and low self-esteem. Body shaming includes a wide range of behaviors, such as fat-shaming, tall-shaming, short-shaming, skinny-shaming, and derogatory comments targeting body dysmorphia, facial or body asymmetry, eczema, and acne. These harmful behaviors affect diverse populations, particularly vulnerable groups such as

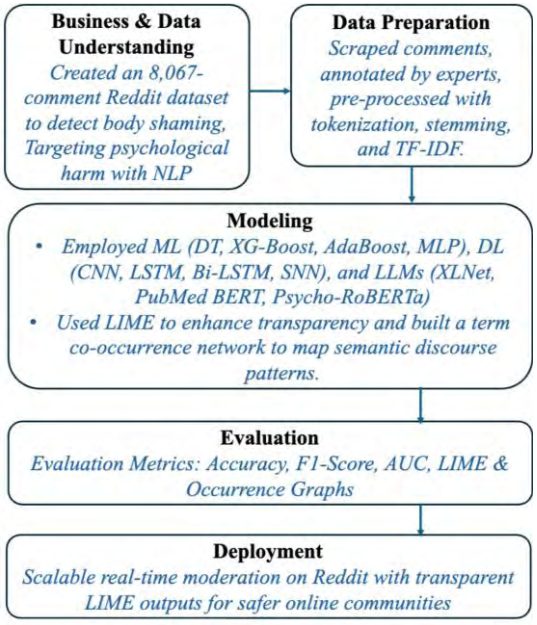


Figure. 1. Proposed Framework for Body Shaming Detection Based on the CRISP-DM Methodology

adolescents and individuals with body image concerns. Notably, body shaming can be externally directed, where users are targeted by peers, strangers, or anonymous individuals, or self-directed, where users internalize societal standards and engage in self-criticism. By analyzing both forms, this study seeks to capture the full spectrum of body shaming discourse and its psychological consequences. Reddit was selected as the primary data source due to its vast and diverse user base, encompassing millions of users who participate in thousands of topic-based subreddits. Its open, community-driven structure encourages candid discussion, making it an ideal platform to investigate the dynamics and impact of body shaming. To support this investigation, we developed a methodological framework focused on technological precision and systematic data collection. Using Octoparse 8, a free and open-source web scraping tool, we gathered a dataset of 8,067 Reddit comments posted between June and November 2024. This dataset includes both body shaming and non-body shaming content, enabling comparative analysis. Body shaming comments were identified using hashtags sourced from existing literature, including #fatshaming, #tallshaming, #shortshaming, #skinnyshaming, #dysmorphia, #asymmetry, #eczema, and #acne, covering a broad spectrum of physical traits and conditions. In contrast, non-body shaming comments were labeled using keywords and hashtags related to other forms of mental distress and cyberbullying, such as depression, suicidal ideation, cursing, offensive language, and hate speech. This dual-labelling approach enables us to more effectively distinguish body shaming discourse from general online toxicity.

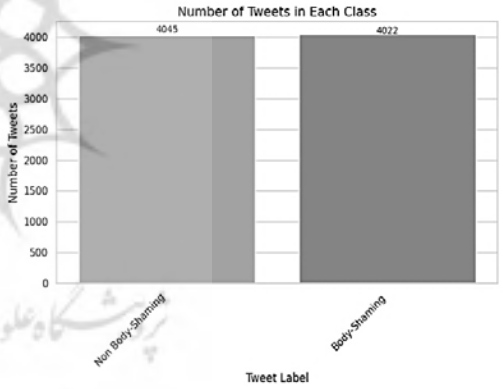
To illustrate the emotional tone and thematic diversity of the comments, Table 1 presents selected samples of Reddit comments along with their corresponding hashtags, illustrating the variety of body shaming expressions captured in our dataset. This provides context for understanding the diversity and nature of the data analyzed.

To ensure the reliability and precision of the dataset, English literature experts trained in information science manually annotated each comment, classifying them into two distinct categories: body shaming, which includes derogatory remarks targeting physical attributes, and non-body shaming, which encompasses general offensive language or content related to mental health. This expert-labeled dataset forms the foundation of a rigorous text preprocessing pipeline, developed to enhance the performance of ML, DL, and transformer-based LLMs. The preprocessing workflow begins by standardizing the textual data, converting all text to lowercase, expanding contractions (e.g., "can't" to "cannot"), and normalizing informal language (e.g., "gonna" to "going to"). It also involves removing extraneous elements such as special characters (%, @, !), punctuation, and numerical digits to eliminate noise and ensure consistency. After cleaning, the text is tokenized into smaller units, typically words or subwords, based on the specific requirements of the modeling technique; for instance, subword tokenization is used for transformer models like BERT. The process continues with normalization techniques, including stemming, which reduces words to their root forms by trimming suffixes (e.g., "shameful" to "shame"), and lemmatization, which uses linguistic rules to derive base word forms with greater grammatical accuracy. To further refine the input, stop words such as "the," "is," and "and," which contribute little to the semantic meaning, are removed, and common spelling errors are corrected to improve overall data quality. Once the text is cleaned and structured, TF-IDF is applied to identify and weight the most informative words, enabling the model to detect important patterns across comments. The structure and composition of the dataset are depicted in Figure 2, where part (a) illustrates the distribution of 8,067 comments across the two categories, body shaming and non-body shaming, highlighting their relative frequency, while part (b) compares the average comment lengths in each category, offering insights into differences in verbosity and expression style.

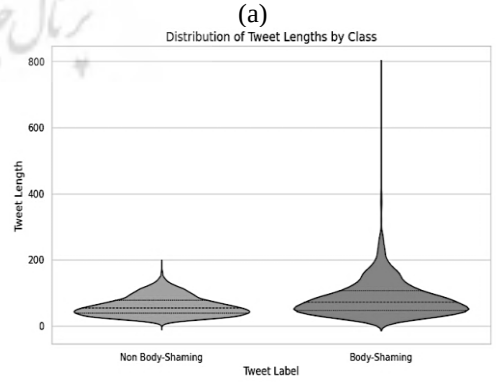
Table 1. Sampled Comments Corresponding to Analyzed Hashtags

# Hashtag	*Content*
#shortshaming	I have many tallll guy friends and they love to always point out when i get there & shorter than them. #shortshaming seriously lame fuuu***
#tallshaming	Fu******nerddddd swaggggggggg black af suff from #tallshaming + skinny ugly hehhe dmn @#%
#fatshaming	I'm pro-#fatshaming in every venue. Put the fork down, tubby.
#skinny shaming	Fuck 'fat shaming' I'm sick of #skinny shaming. Sorrrrrry my metabolism is fly even after having a child :(((@%
#dismorphia	I was so lean back then, also a fact - I suffer on body #dysmorphia and I used to hate my body being overly lean.
#assymetric	I'm insecure about my double eyelids bc they're #assymetric af plus disgusting Oops!
#acne	For mo & be sooo disappointed and insecure about ma face at that time I lost family I lost a lot of hope and I was going through a hard time and my #acne got so out of control……!
#egzema	#egzema is backkkk gonna cry mann%#!

This dataset serves as the foundation for a detailed investigation into the prevalence and nuances of body shaming discourse within online communities, enabling robust model development and evaluation.



(a)



(b)

Figure. 2. Dataset Composition: (a) Distribution of 8,067 Comments Across Body Shaming and Non-Body Shaming Categories, (b) Comparison of Average Comment Lengths by Category

3.2. Model Selection and Training

This study employs a range of ML, DL, and LLMs for body shaming detection. ML models include Decision Trees (DT), Extreme Gradient Boosting (XG-Boost), Adaptive Boosting (AdaBoost), and Multilayer Perceptron (MLP). DL models comprise CNN, LSTM, Bi-LSTM, and Spiking Neural Networks (SNN). LLMs include BERT, RoBERTa, and XLNet. Table 2 summarizes the different ML, DL, and LLMs employed in this study for body shaming detection. The table details each model's architecture, description, and particular strengths relevant to the classification task.

To demonstrate the advanced language models adopted in this study, we focus on BERT, RoBERTa, and XLNet, which represent state-of-the-art developments in NLP. These models are widely acclaimed for their ability to understand nuanced linguistic patterns and capture complex contextual relationships within text. Built on the transformer architecture, BERT and RoBERTa utilize multi-layer encoder structures with self-attention mechanisms, enabling the models to weigh the importance of different words within a sentence, effectively capturing inter-word dependencies and semantic relationships. This self-attention capability is crucial for identifying subtle cues and contextual variations in body shaming discourse. Figure 3 illustrates the architectural designs of BERT and RoBERTa, highlighting their structural parallels and key advancements. As shown, both models utilize deep encoder layers that process input sequences concurrently, enabling efficient and context-sensitive representation learning. Their robust architectures, paired with extensive pretraining on large datasets, make them highly effective for classifying complex and emotionally charged social media content, such as body shaming.

Table 2. Overview of Models Used for Body Shaming Detection

Categoty (Model)	*Description*
ML (XG-Boost)	Ensemble method boosting weak learners for high predictive accuracy [21]
ML (Ada-Boost)	Boosting algorithm improving performance on complex datasets [21]
ML (MLP)	Neural network capturing nonlinear relationships in text data [22]
ML (DT)	Simple decision rules for interpretable classification [22]
DL (CNN)	Captures spatial text hierarchies, ideal for text classification [23]
DL (LSTM)	Models long-term dependencies for sequential text analysis [23]
DL (Bi-LSTM)	Incorporates past and future context for enhanced sentiment analysis [23]
DL (SNN)	Processes data with unique spiking mechanisms for efficiency [24]
LLM (BERT)	Bidirectional transformer capturing contextual nuances [25]
LLM (RoBERTa)	Optimized BERT with larger training corpus for improved performance [25]
LLM (XLNet)	Autoregressive model with permutation-based context learning [25]

This design enhances their contextual understanding and ability to detect linguistic nuances. BERT's self-attention mechanism enables simultaneous attention to multiple input aspects, while RoBERTa builds on this capability with a larger training corpus and an optimized pre-training methodology, further improving performance. The study also explores domain-specific variants of these models: PubMed BERT and Psycho RoBERTa.

PubMed BERT is fine-tuned on the extensive biomedical literature in the PubMed database, leveraging specialized knowledge to excel in tasks such as biomedical text classification and named trained on a comprehensive corpus of psychological

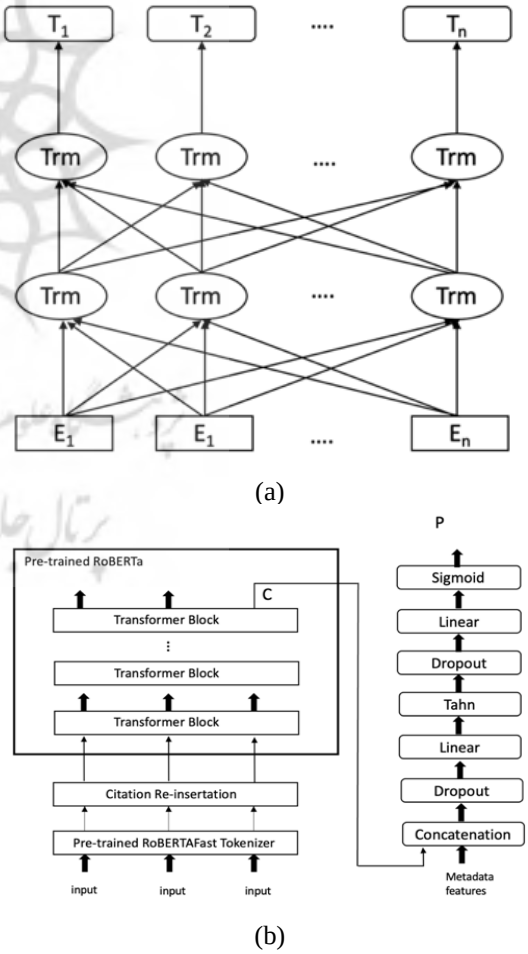


Figure. 3. Architecture of (a) BERT and (b) RoBERTa Transformer Models [25]

and social sciences literature, demonstrates superior entity recognition. Psycho RoBERTa, pre-performance in NLP tasks related to psychology and mental health research. By incorporating these domain-specific models, the study achieves greater accuracy and context sensitivity in processing biomedical and psychological data, significantly enhancing the effectiveness of NLP tasks in this research. This study also explores XLNet, an advanced language model that processes text in a flexible order, unlike traditional models that assume words are independent. XLNet uses Permutation Language Modeling, which considers all possible word sequences to capture context from both directions (e.g., understanding "big fat" as a phrase, not isolated words). It employs two-stream self-attention: a content stream that analyzes word meanings and a query stream that focuses on word positions, enhancing its ability to detect nuanced body shaming. Masked attention generates permutations, allowing XLNet to maintain bidirectionality. The two-stream mechanism addresses the challenge of the query stream's limited information, enhancing target-oriented context awareness. The study further investigates fine-tuned and partially tuned variants of PubMed BERT, XLNet, and Psycho-RoBERTa for domain-specific text classification. Fine-tuning updates all model layers, including the transformer encoder and classifier head, capturing complex patterns in biomedical and psychological texts. In contrast, partial tuning freezes the encoder and trains only the classifier head, reducing computation while leveraging learned representations. This approach is useful when resources are constrained or the model is already domain-aligned. Both approaches optimize a loss function; the loss function used during training is the categorical cross-entropy, defined as Equation (1) [26]:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(y_{i,c}\right) \qquad (1)$$

Where N is the total number of samples, C is the number of classes, $y_{i,c}$ is the true label for class C of sample i (1 if correct, 0 otherwise), and $y_{i,c}$ is the predicted probability for class C of sample i.

During fine-tuning, all model parameters θ, including both the transformer weights and the classifier head, are updated via backpropagation to minimize the loss LL. In partial tuning**,** only the classifier parameters θ_{cls} are updated, while the transformer parameters θ_{trans} remain frozen. PubMed-BERT and XLNet, pre-trained on large biomedical corpora, were fine-tuned to adapt to the specific language of medical texts. Psycho-RoBERTa, tailored for psychological domains, was both fine-tuned and partially tuned to balance classification performance with computational efficiency [25-26]. This dual strategy ensured accurate and efficient feature extraction across complex biomedical and psychological datasets.

3.3. Model Interpretability

This study employs LIME to enhance the interpretability of machine learning predictions, particularly for complex models like transformer-based LLMs, where decision-making processes are often opaque. LIME improves model transparency by identifying key text features that influence predictions [27]. For example, in a comment like "You're too fat to wear that," LIME might highlight "fat" as a primary indicator of body shaming, assigning it a high weight in the model's decision. In content moderation, this approach reveals critical features driving classifications, helping users understand prediction rationales. Figure 4 illustrates LIME's mechanism in explaining model outputs.

The Figure above illustrates the basic concept behind LIME. In this diagram, the red cross marks the specific data point we want to explain. The background color shows the decision boundary of a complex ML: the pink area represents one class (such as "positive"), and the blue area represents another class (such as "negative"). The shape of these regions indicates that the model being explained is highly complex and nonlinear. To generate an explanation, LIME starts by creating many new samples around the original data point (the red cross). These nearby points are slightly altered versions of the original and are shown in the image as small red pluses and blue circles. Each of these new points is passed through the original model to get a prediction, which helps LIME understand how the model behaves in the local area around the instance of interest. LIME then trains a simple and interpretable model, such as a linear model, using only these nearby samples. This local model is shown as the dashed line in the image. Although this dashed line does not match the complex global shape of the original model's decision boundary, it does a good job of approximating the model's behavior in the local neighborhood of the red cross. This approach allows LIME to offer an explanation that is both accurate for that small region and easy to

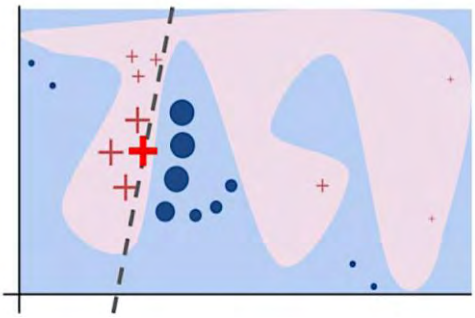


Figure. 4. Illustration of the LIME Mechanism [27]

understand. The key idea is that we don't need to explain the entire model at once. Instead, we focus on explaining one prediction at a time by zooming in on a small area around the input and fitting a simple model there. LIME is also model-agnostic, meaning it can be used with any MLs, no matter how complex it is. By providing explanations that are both faithful to the model locally and understandable to humans, LIME helps users build trust in AI systems. However, it is worth noting that because LIME only looks at a small part of the model's behavior, its explanations may not always reflect the full picture, especially in cases where the model behaves unpredictably or inconsistently in nearby regions [27-28].

3.4. Topic Modeling by Term Co-occurrence Matrix (Graph Visualization)

To quantify the semantic proximity between terms, we computed a term co-occurrence matrix by first generating a Document-Term Matrix (DTM) using the bag-of-words model. In this model, each document from the corpus is converted into a vector where each element represents the frequency of a specific term in that document. This results in a DTM where rows correspond to documents and columns represent terms. The co-occurrence matrix was then calculated by taking the dot product of the DTM's transpose with itself, expressed as Equation (2) [29]:

$$CoMatrix = DTM^T \times DTM \quad (2)$$

This operation yields a symmetric matrix where each entry *(i,j)* reflects how many documents contain both term *i* and term j, effectively capturing their co-occurrence frequency within the same context. In other words, a high value in the *(i,j)* entry indicates that the two terms often appear together in the same documents, suggesting a stronger semantic relationship.

To build the network, we retained only the upper-triangular entries with non-zero values, thereby avoiding duplicate edges (since the matrix is symmetric) and self-loops (where a term would be connected to itself). This matrix formed the basis of an undirected weighted graph, where nodes represent terms and edge weights represent the strength of their co-occurrence. In this constructed graph, more frequently co-occurring term pairs result in heavier (stronger) connections, which is useful for analyzing semantic structures or visualizing relationships between terms in the corpus.

4. Results

This study evaluated the performance of traditional MLs, DLs, and transformer-based LLMs in detecting body shaming within a novel dataset of 8,067 Reddit comments collected from June to November 2024. Fine-tuned Psycho-RoBERTa outperformed traditional ML and DL models in detecting body shaming, as shown in Table 3. The accuracy of 0.98 reflects its ability to capture nuanced, context-dependent discourse. Table 3 compares the performance of various classification models on the body shaming detection task. It includes key evaluation metrics such as Accuracy, F1-score, and AUC to highlight the effectiveness of traditional ML, DL, and transformer-based LLMs. It is important to note that in the table below, "P" denotes partially fine-tuned LLMs, while "F" indicates fully fine-tuned LLMs.

The results underscore distinct performance trends among various machine learning, deep learning, and language model architectures. Traditional machine learning methods like XG-Boost and AdaBoost achieved solid accuracy scores (0.972), reflecting their reliability with structured data, yet they lagged in fully capturing nuanced semantic relationships due to their dependence on simpler, manually engineered features. Deep learning models such as CNN, Bi-LSTM, and SNN demonstrated improved accuracy levels (ranging from 0.97 to nearly 0.98), leveraging their capacity for hierarchical feature extraction and sequential data

Table 3. Performance Metrics for Traditional ML, DL, and Transformer-Based LLMs in Body Shaming Detection

Model	*Accuracy*	*F1-Score*	*AUC*
DT	0.95530	0.96504	0.97530
XG-Boost	0.97211	0.97206	0.99753
AdaBoost	0.97211	0.97227	0.99578
MLP	0.97273	0.97232	0.99590
LSTM	0.49628	0.49069	0.50062
CNN	0.97955	0.97931	0.99735
Bi-LSTM	0.97397	0.97413	0.99483
SNN	0.97583	0.97554	0.99680
P-XLNet	0.49628	0.00	0.5
P-PubMed Bert	0.72056	0.63241	0.76962
P-Psycho-Roberta	0.96224	0.95162	0.99523
F-PubMed Bert	0.96822	0.98833	0.99654
F-XLNet	0.97007	0.98678	0.99906
F- Psycho-Roberta	0.98186	0.99406	0.99018

modeling, which enabled them to better grasp complex patterns within the dataset.

In contrast, standard LSTM struggled significantly, achieving markedly poor performance (accuracy: 0.496). This highlights its limitations when processing linguistically and contextually rich text, especially when domain adaptation is insufficient. Similarly, partially tuned transformer-based models like P-XLNet also performed weakly, emphasizing the importance of task-specific tuning in extracting value from state-of-the-art architectures.

Fine-tuned LLMs such as F-PubMed Bert, F-XLNet, and F-Psycho-Roberta outperformed other approaches, with the highest accuracy and F1-scores (up to 0.981 for accuracy and 0.994 for F1-score). This can be attributed to their advanced contextual understanding and domain-adapted representations, which allow for the recognition of subtle cues and linguistic complexity that simpler models miss. In particular, the combination of Psycho-RoBERTa's psychological pre-training and XLNet's permutation-based learning yielded strong detection of intricate language patterns associated with body shaming.

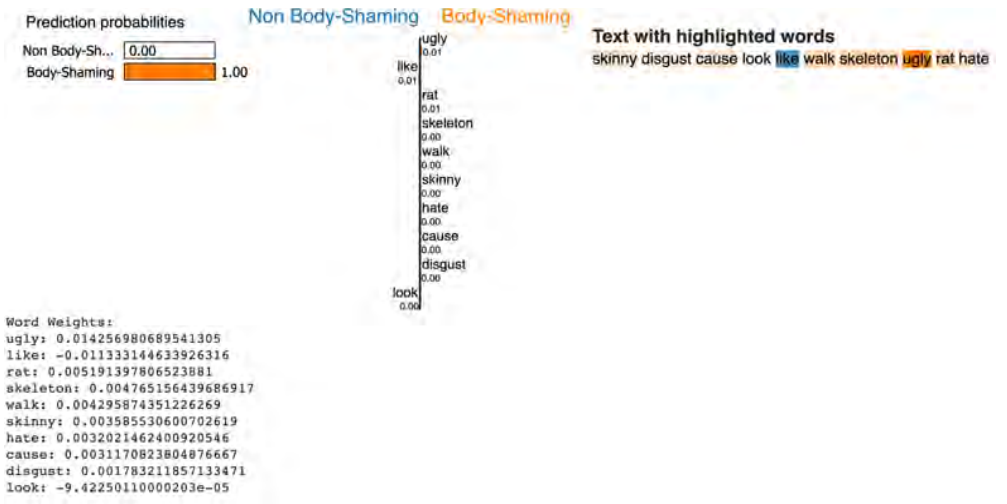
LIME enhanced interpretability by identifying critical input features that influenced the models' predictions. Notably, terms such as "fat," "ugly," "skinny," and "acne" emerged as key predictors across all models, with particularly strong importance in fine-tuned architectures. These findings elucidate the models' decision-making pathways and reveal the linguistic patterns most strongly associated with body shaming, thereby supporting transparency and fairness assessments. Figure 5 presents the models' prediction probabilities for distinguishing body-shaming content, alongside the most influential keywords and their corresponding weights.

These terms align with the derogatory and emotionally charged language prevalent in body shaming discourse. For instance, in Psycho-RoBERTa's predictions, "fat" and "ugly" were assigned high weights, reflecting their strong influence on identifying harmful content. Conversely, partially tuned models like Partial XLNet showed inconsistent feature importance, underscoring their limited effectiveness. The LIME outputs, visualized in Figure 5, reveal distinct patterns: fine-tuned models (a, b, c) consistently prioritized contextually relevant terms, while partially tuned models (d, e, f) exhibited scattered or negligible feature contributions, explaining their lower performance. To enhance the interpretability of this research, Figure 6 introduces a novel topic modelling technique using a Co-occurrence Term Matrix to visualize semantic relationships in body shaming discourse.

The term co-occurrence network, constructed using a DTM and visualized in Figure 6, reveals semantic relationships within the dataset. The co-occurrence matrix was computed as the dot product of the DTM's transpose (Equation (2)), creating an undirected weighted graph where nodes represent terms and edges denote co-occurrence frequencies. The graph, generated using networkx and matplotlib, positions central nodes like "shame," "body," and "fat" at the core, indicating their frequent co-occurrence with other terms. Peripheral nodes, such as "asymmetry" and "eczema," reflect less frequent but contextually significant themes. Clusters around terms like "depression" and "anxiety" highlight the linkage between body shaming and mental health, aligning with prior research on psychological harm [9, 29].

5. Discussion

This study demonstrates the superior performance of fine-tuned Psycho-RoBERTa (accuracy: 0.98, F1-score: 0.994, AUC: 0.990) in detecting body shaming on Reddit, surpassing traditional MLs like DT (accuracy: 0.955) and DL models like LSTM (accuracy: 0.496). The transformer architecture of LLMs, leveraging self-attention mechanisms, excels at capturing nuanced linguistic patterns, such as implicit insults or emotionally charged terms prevalent in Reddit's unfiltered, subreddit-driven discourse. Psycho-RoBERTa, pre-trained on psychological literature, shows exceptional sensitivity to terms like "dysmorphia" and "anxiety," aligning with the dataset's focus on mental health-related content [7-9]. In contrast, prior studies on platforms like X and YouTube [14-15] reported lower accuracies (e.g., 0.94 with ALBERTO on Instagram [19]) due to smaller datasets and less context-aware models. XLNet's permutation-based learning further enhances detection by adeptly handling Reddit's fragmented, context-shifting comments. Unlike traditional ML models (e.g., DT, SVM), which rely on static feature extraction like TF-IDF, and sequential DL models like LSTM, which struggle with long-range dependencies in complex social media texts, transformers capture bidirectional context, enabling robust detection of harmful content. This performance gap highlights the transformative role of transformer-based models in addressing the complexities of social media harassment. The term co-occurrence network graph (Figure 6) provides a novel lens into the semantic structure of body shaming discourse. By computing the co-occurrence matrix as the dot product of the DTM's transpose, the study constructs a weighted graph where nodes represent terms and edges reflect their co-occurrence frequency. Central nodes like "shame," "body," and "fat" dominate, indicating their frequent use in harmful comments, while clusters linking "fat" to "weight" and "depression" corroborate prior research on body shaming's psychological impact [9].


Prediction probabilities
Non Body-Sh... 0.00
Body-Shaming 1.00
Non Body-Shaming
Body-Shaming
ugly 0.01
like 0.01
rat 0.01
skeleton 0.00
walk 0.00
skinny 0.00
hate 0.00
cause 0.00
disgust 0.00
look 0.00
Text with highlighted words
skinny disgust cause look like walk skeleton ugly rat hate
Word Weights:
ugly: 0.014256980689541305
like: -0.011333144633926316
rat: 0.005191397806523881
skeleton: 0.004765156439686917
walk: 0.004295874351226269
skinny: 0.003585530600702619
hate: 0.0032021462400920546
cause: 0.0031170823804876667
disgust: 0.0017832118571334 71
look: -9.42250110000203e-05

(a)


Prediction probabilities
Non Body-Sh... 0.00
Body-Shaming 1.00
Non Body-Shaming
Body-Shaming
like 0.01
ugly 0.01
skinny 0.01
look 0.01
rat 0.00
skeleton 0.00
hate 0.00
disgust 0.00
walk 0.00
cause 0.00
Text with highlighted words
skinny disgust cause look like walk skeleton ugly rat hate
Word Weights:
like: -0.009938567940339179
ugly: 0.00828371229439659
skinny: 0.007072179632415 22
look: -0.005535294193266718
rat: 0.004712924335026901
skeleton: 0.004382575629631271
hate: 0.003578514934793321
disgust: 0.003352963093690521
walk: 0.0033040718740433165
cause: 0.0029604042838646534

(b)


Prediction probabilities
Non Body-Sh... 0.00
Body-Shaming 1.00
Non Body-Shaming
Body-Shaming
ugly 0.00
hate 0.00
like 0.00
disgust 0.00
skinny 0.00
look 0.00
rat 0.00
walk 0.00
skeleton 0.00
cause 0.00
Text with highlighted words
skinny disgust cause look like walk skeleton ugly rat hate
Word Weights:
ugly: 9.766931195433055e-05
hate: 8.089050255897955e-05
like: -8.037843613479142e-05
disgust: 6.605004355534 72e-05
skinny: 5.5448028025040910-05
look: -4.7977538300922234e-05
rat: -2.498316758069626e-05
walk: -1.0330542647875483e-05
skeleton: -9.272543156692197e-06
cause: 4.804435536918142 5e-06

(c)

(d)

(e)

(f)

Figure. 5. LIME Visualizations for Body Shaming Classification: (a) Fine-Tuned PubMed BERT, (b) Fine-Tuned XLNet, (c) Fine-Tuned Psycho-RoBERTa, (d) Partially Tuned XLNet, (e) Partially Tuned Psycho-RoBERTa, (f) Partially Tuned PubMed-BERT

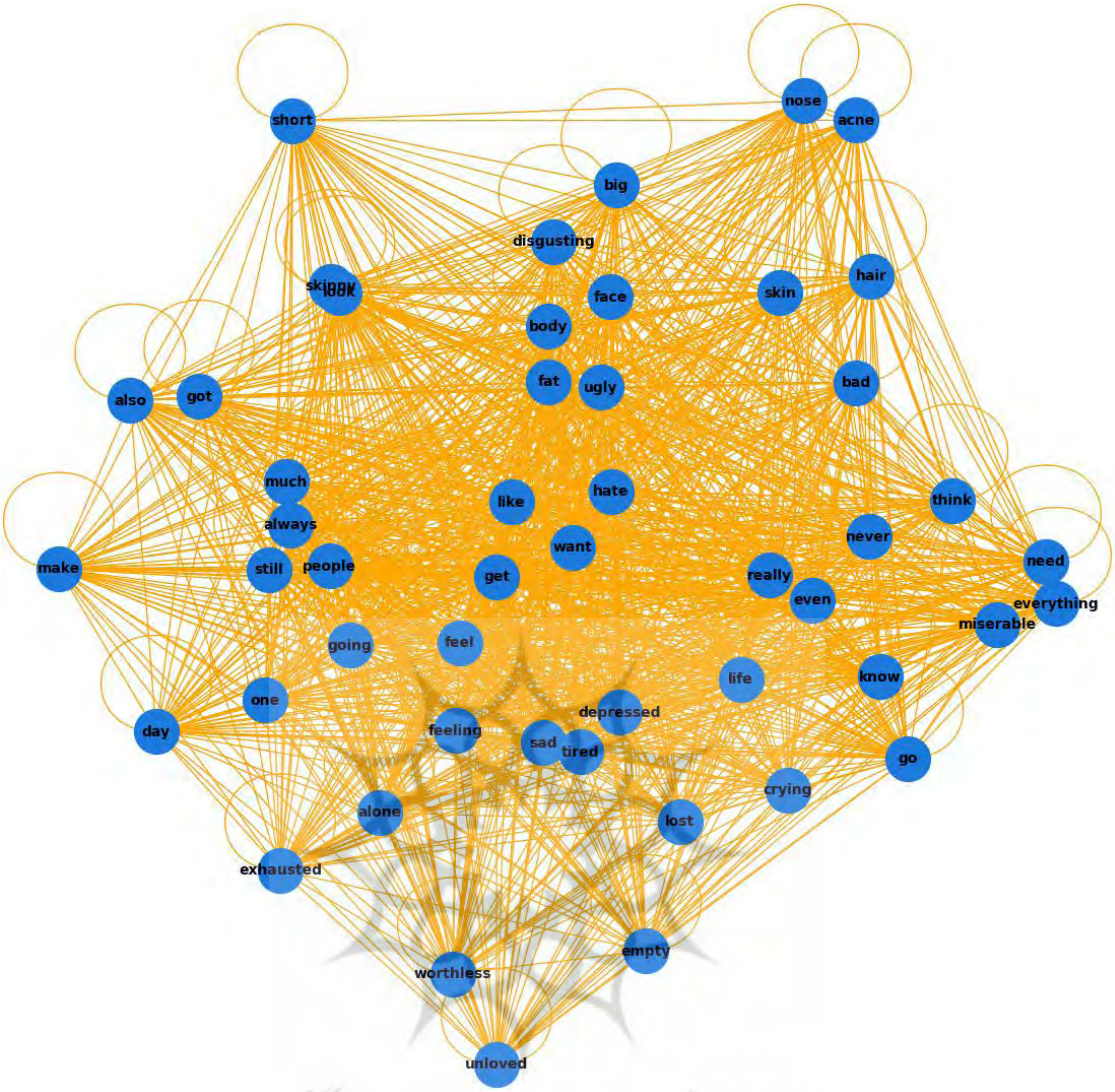


Figure. 6. Term Co-occurrence Network Graph for Body Shaming Preprocessed Discourse

Peripheral nodes, such as "eczema" and "asymmetry," reveal niche but significant themes, highlighting the diversity of body shaming expressions. Visualized using the spring_layout algorithm in NetworkX, the graph positions densely connected terms centrally, forming thematic clusters that map the discourse's structure. This approach not only validates the models' focus on key terms but also offers insights into the thematic patterns of harmful content, supporting targeted moderation strategies. The LIME enhance the study's impact by providing transparency into model decisions, fostering trust in AI-driven content moderation. LIME identifies key predictors like "fat," "ugly," and "acne," which consistently drive body shaming classifications in fine-tuned models (Figure 5). These terms align closely with the co-occurrence graph's central nodes, reinforcing their role as core shaming indicators. For example, Psycho-RoBERTa's high weighting of "fat" mirrors the graph's clustering, confirming its prominence in harmful discourse. However, LIME's focus on local explanations limits its ability to capture broader patterns, such as the interplay between rare terms like "eczema" and "insecure." A proposed segregated discussion tree could complement LIME by modeling hierarchical term relationships, illustrating how "shame" branches into subthemes like "body" to "fat" to "depression," offering a richer understanding of discourse dynamics. Combining LIME with such approaches could balance local and global interpretability, ensuring both transparency and comprehensive insights. The study's dataset of 8,067 Reddit comments addresses a critical gap in cyberbullying research, as prior studies on platforms like X or YouTube often relied on smaller, less representative datasets (<2,000 samples). By capturing Reddit's anonymity-driven, subreddit-specific discourse, the dataset encompasses both

externally directed shaming (e.g., "Put the fork down, tubby") and self-directed shaming (e.g., "I hate my body being overly lean"), broadening the scope of analysis. This dual focus highlights the complexity of body shaming and the need for nuanced detection approaches. However, the hashtag-based collection may miss subtle or implicit shaming, suggesting opportunities for unsupervised methods like clustering or few-shot learning to enhance sensitivity to nuanced cues. Together, the dataset, advanced LLMs, and interpretability tools establish a robust framework for tackling body shaming, advancing the field of NLP and online harassment detection.

5.1. Theoretical Contributions

Theoretical contributions enrich NLP and body shaming research. The dataset advances platform-specific resources, addressing Reddit's unique dynamics. The co-occurrence network deepens understanding of semantic structures, supporting theories of psychological harm [29]. The comparative analysis of ML, DL, and LLMs reinforces the paradigm shift toward transformers, while domain-specific models like Psycho-RoBERTa highlight the value of specialized pre-training. However, challenges remain. Partial tuning (e.g., Partial XLNet at 0.496 accuracy) underperformed, underscoring the need for full fine-tuning in complex tasks. Lightweight LLMs like DistilBERT could balance efficiency and performance, especially for smaller platforms, while GPT-based models could enhance the detection of implicit shaming. Longitudinal studies and multi-platform datasets are critical to ensure models evolve with social media's dynamic nature, maintaining relevance in combating harassment.

5.2. Practical Implications

The practical implications of this study are substantial for social media platforms aiming to combat body shaming. The high accuracy of Psycho-RoBERTa (0.98) enables real-time content moderation with minimal false positives, significantly reducing the burden on human moderators. For instance, Reddit could integrate Psycho-RoBERTa into its AutoModerator system to flag comments containing terms like 'fat' or 'ugly' in high-risk subreddits like r/AmITheAsshole, where body shaming is prevalent. The term co-occurrence graph, identifying clusters like 'depression' and 'fat,' allows platforms to proactively target subreddits for stricter guidelines or embed mental health resources, fostering inclusivity. To scale these models for Reddit's massive comment volume, platforms could leverage cloud-based GPU clusters or adopt lightweight LLMs like DistilBERT, balancing accuracy with computational efficiency. LIME's transparency ensures fair moderation by providing clear rationales for flagging decisions, crucial for user trust and appeal processes. However, ethical considerations, such as minimizing false positives to avoid unfair penalties and ensuring data collection complies with privacy regulations like GDPR, are critical for responsible implementation. Public awareness campaigns can leverage the graph's insights to educate users about body shaming's psychological impact, promoting empathetic online interactions and reducing harmful discourse

5.3. Limitations and Future Directions

The study's limitations highlight opportunities for future research to enhance body shaming detection. The Reddit-specific dataset, while rich, may not generalize to platforms like X or TikTok, where discourse patterns differ due to real-time interactions or visual content, respectively. This limits the model's applicability to diverse social media ecosystems. The hashtag-based collection, while effective for explicit shaming (e.g., #fatshaming), may underrepresent implicit or sarcastic remarks, reducing sensitivity to subtle harassment. The binary classification framework oversimplifies the complexity of online harassment, missing overlapping behaviors like cyberbullying and hate speech. Future work could address these by integrating multi-platform datasets to capture platform-specific dynamics, such as comparing Reddit's anonymity-driven discourse with TikTok's visual-centric content. Advanced techniques, like GPT-4's few-shot learning, could detect implicit shaming by generalizing from limited examples, while multi-label models could classify comments across multiple harassment categories. Integrating SHAP or BERTopic could complement LIME by capturing global feature contributions or latent topics, respectively. The computational cost of fine-tuning LLMs may limit adoption, suggesting exploration of lightweight models like DistilBERT. Longitudinal studies and graph-based community detection, analyzing user interaction networks to identify toxic subreddits, could further enhance dynamic moderation strategies, ensuring models evolve with social media's rapidly changing landscape.

6. Conclusions

This research advances body shaming detection on Reddit by introducing a dataset of 8,067 comments from June to November 2024, capturing the platform's unique discourse. Fine-tuned Psycho-RoBERTa excels (accuracy: 0.98) in identifying harmful content, outperforming traditional MLs and DLs due to its psychological pre-training. Transparent moderation is achieved through LIME, which highlights influential terms like "fat," while a term co-occurrence graph reveals connections between "shame" and "depression," informing mental health-focused interventions. This framework offers scalable tools for real-time content flagging and promotes inclusive online spaces. Future work

should incorporate multi-platform data, multi-label models, and efficient LLMs like DistilBERT to address implicit shaming and enhance applicability. By leveraging NLP, this study fosters safer social media environments.

Declaration

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-pro]fit sectors.

Author's Contribution

S.T.: Study design, data acquisition and preparation, technical deployment, result evaluation, full manuscript preparation, and addressing reviewer comments.

N.A.: Supervision, conceptualization, data validation, critical revision of manuscript, guidance on study design and interpretation of results, and advising on addressing reviewers' comments.

Conflict of interest

The authors declare that no conflicts of interest exist.

Refrences

[1] A. G. Philipo, D. Sebastian Sarwatt, J. Ding, M. Daneshmand, H. Ning, "Assessing Text Classification Methods for Cyberbullying Detection on Social Media Platforms, " *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 7602-7616, 2025. https://doi.org/10.1109/TIFS.2025.3588728.

[2] T. Prince, K. E. Mulgrew, C. Driver, et al., "The body image related cyberbullying picture series (BRC-PicS): developed for use in research relating to cyberbullying, body image and eating disorders among female adolescents, " *Curr. Psychol*., vol. 44, pp. 2747-2760, 2025. https://doi.org/10.1007/s12144-025-07316-x.

[3] R. Wang, B. Ye, P. Wang, "Appearance comparison on social networking sites and body shame: The role of negative body talk and perceived sociocultural influences on body image, " *J. Health Psychol*., vol. 30, no. 2, pp. 224-237, 2024. https://doi.org/10.1177/13591053241245100.

[4] B. Bonthu, P. Abhay, L. S. Gottipati and G. K. Vamsi, "CivilityCheck: An Integrated Natural Language Processing and Machine Learning Framework to Detect Hateful and Offensive Language, " *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2024, pp. 985-988. https://doi.org/10.1109/ICSCSS60660.2024.10625318.

[5] M. H. Obaida, S. M. Elkaffas, S. K. Guirguis, "Deep Learning Algorithms for Cyber-Bulling Detection in Social Media Platforms", *IEEE* Access, vol. 12, pp. 76901-76908, 2024. https://doi.org/10.1109/ACCESS.2024.3406595.

[6] S. Talebi, N. Abdolvand, "Leveraging Explainable Artificial Intelligence for Comparative Large Language Models in Body Shaming Detection," *International Conference on Web Research (ICWR)*, Tehran, Iran, 2025, pp. 1-6, https://doi.org/10.1109/ICWR65219.2025.11006221

[7] E. Corradini, "The dark threads that weave the web of shame: A network science-inspired analysis of body shaming on Reddit, " *Information*, vol. 14, no. 8, p. 436, 2023. https://doi.org/10.3390/info14080436.

[8] J. S. Jacob, N. Panwar, "Effect of age and gender on dietary patterns, mindful eating, body image and confidence," *BMC Psychol*., vol. 11, p. 264, 2023. https://doi.org/10.1186/s40359-023-01290-4.

[9] J. M. Cullin, "Biological normalcy and body fat: Obesity prevalence, fat stigma, and allostatic load among late adolescents and young adults, " *Amer. J. Biol. Anthropol.*, vol. 181, no. 4, pp. 575-587, 2023. https://doi.org/10.1002/ajpa.24752.

[10] S. F. Nurul Fitri H, F. Fattah, H. Azis, "Comparative analysis of machine learning algorithm variations in classifying body shaming topics on social media X," *Int. J. Data Sci. (IJODAS)*, vol. 5, no. 2, pp. 121-131, 2024. https://doi.org/10.56705/ijodas.v5i2.82.

[11] R. Puvanendran et al., "Comparative Analysis of Resampling Techniques on Class Imbalance in Body Shaming Phrase Detection," *Moratuwa Engineering Research Conference (MERCon)*, Moratuwa, Sri Lanka, 2024, pp. 49-54. https://doi.org/10.1109/MERCon63886.2024.10688774.

[12] X. Li, R. Li, and Y. Zheng, "Body shame Twitter movement text mining and data analysis by applying MDCOR," *Lect. Notes Educ. Psychol. Public Media*, vol. 10, pp. 7-13, 2023. https://doi.org/10.54254/2753-7048/10/20230013.

[13] J. Haerul Jaman, J. Jajam, and H. Hannie, "Sentiment analysis of the body-shaming beauty vlog comments," in *Proc. EAI Int. Conf. Smart Comput. Commun*, 2019. https://doi.org/10.4108/eai.12-10-2019.2296530.

[14] S. F. Nurul Fitri H, F. Fattah, H. Azis, "Comparative Analysis of Machine Learning Algorithm Variations in Classifying Body Shaming Topics on Social Media X", *Ijodas*, vol. 5, no. 2, pp. 121-132, 2024. https://doi.org/10.56705/ijodas.v5i2.82.

[15] V. Reddy, H. Abburi, "You Are Big, S/he Is Small' Detecting body shaming in online user content," *SocInfo*, vol. 13618, pp. 389-397, 2022. https://doi.org/10.1007/978-3-031-19097-1_25.

[16] S. Mundra, N. Mittal, R. Nayak, "Prototypical network based few shot learning to detect Hindi–English code-mixed offensive text," Social Network Analysis and Mining, vol. 15, no. 49, 2025. https://doi.org/10.1007/s13278-025-01431-0.

[17] E. Y. Daraghmi, S. Qadan, Y. A. Daraghmi, R. Yousuf, O. Cheikhrouhou, M. Baz, "From Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection, " *IEEE Access*, vol. 12, pp. 103504-103519, 2024. https://doi.org/10.1109/ACCESS.2024.3431939.

[18] E. Halim, E. Mardiah, H. L. Sunarta, R. A. Putri, "From mirrors to mindsets: The chain reaction of social media, body shame, and body image perceptions," *024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, 2024, pp. 647-652. https://doi.org/10.1109/ISRITI64779.2024.10963624.

[19] F. Grasso, A. Valese, and M. Micheli, "Body-shaming detection and classification in Italian social media," *Natural Language Processing and Information Systems*, vol. 14762, pp. 256-270, 2024. https://doi.org/10.1007/978-3-031-70239-6_18

[20] T. G. Wijaya, O. N. Pratiwi, I. Darmawan, "Implementation of CRISP-DM to Predict Student Graduation on Time Using Naïve Bayes Algorithm," *International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, Bali, Indonesia, 2024, pp. 233-238, https://doi.org/10.1109/ICICYTA64807.2024.10913067.

[21] I. G. A. N. Lestari, N. M. R. M. Dewi, K. G. Meiliana, I. K. A. A. Aryanto, "Effectiveness of AdaBoost and XGBoost algorithms in sentiment analysis of movie reviews," *JAIC*, vol. 9, no. 2, pp. 258-264, 2025, https://doi.org/10.30871/jaic.v9i2.9077.

[22] B. L. Thanh Thai, T. Takii, H. Tanaka, "Password Classification Using Machine Learning and Natural Language Processing Techniques: Methods and Evaluations," *2024 8th Cyber Security in Networking Conference (CSNet)*, Paris, France, 2024, pp. 147-150, https://doi.org/10.1109/CSNet64211.2024.10851759.

[23] A. S. Alhanaf, M. Farsadi, H. H. Balik, "Fault Detection and Classification in Ring Power System With DG Penetration Using Hybrid CNN-LSTM," *IEEE Access*, vol. 12, pp. 59953-59975, 2024, https://doi.org/10.1109/ACCESS.2024.3394166.

[24] S. S. Park, Y. S. Choi, "Spiking neural networks for physiological and speech signals: A review," Biomed. Eng. Lett, vol. 14, no. 5, pp. 943-954, 2024. https://doi.org/10.1007/s13534-024-00404-0.

[25] Y. R. Devi, A. Bharthepudi, A. Govindarajula, "A Review on Sentiment Analysis Using Transformers and Ensemble methods," *International Conference on Recent Advances in Information Technology (RAIT)*, Dhanbad, India, 2025, pp. 1-6, https://doi.org/10.1109/RAIT65068.2025.11089332.

[26] N. Ding, M. A. Zarrabian and P. Sadeghi, "A Cross Entropy Interpretation of Renyi Entropy for α -leakage," *International Symposium on Information Theory (ISIT)*, Athens, Greece, 2024, pp. 2760-2765, https://doi.org/10.1109/ISIT57864.2024.10619672.

[27] M. Z. Mahmud, M. S. Reza, S. R. Alve, S. Islam, and N. Fahmid, "Advance transfer learning approach for identification of multiclass skin disease with LIME explainable AI technique," *2024 27th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2024, pp. 109-114, https://doi.org/10.1109/ICCIT64611.2024.11021908.

[28] T. A. A. Abdullah, et al., "Sig-Lime: A signal-based enhancement of Lime explanation technique," *IEEE Access*, vol. 12, 2024. https://doi.org/10.1109/ACCESS.2024.3384277.

[29] S. N. S. Al Subhi, A. R. Mikler, M. M. Kubek, "Constructing Co-Occurrence Graphs and Deriving Flood Ontologies for Enhanced Understanding," *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, Bangalore, India, 2024, pp. 1-8, https://doi.org/10.1109/ICICI62254.2024.00009.

Sajedeh Talebi holds a Master's degree in Information Technology Management from Alzahra University. Her research spans Data Mining, Computer Vision, Natural Language Processing, and Human-Computer Interaction, with a focus on the impact of emerging technologies, such as explainable AI, reasoning AI, and diffusion models, on computer vision and language algorithms. She also applies these technologies to marketing, social media, and medical domains.

Dr.Neda Abdolvand is an Associate Professor of Information Technology at Alzahra University, specializing in the application of artificial intelligence and machine learning to sectors such as social media analysis, marketing, auditing, healthcare, and sustainability. Her research integrates advanced analytics and intelligent systems to solve complex, real-world problems, with numerous publications in high-impact journals. She has led industry–academic collaborations, supervised over 50 theses.