

# Body-Shaming Detection and Classification in Italian Social Media

Francesca Grasso<sup>\*[0000-0001-8473-9491]</sup>, Alberto Valesè<sup>[0009-0006-7221-5061]</sup>, and  
Marta Micheli<sup>[0009-0003-4562-0334]</sup>

Department of Computer Science, University of Turin  
Corso Svizzera 185 - 10149 Turin, Italy  
{fr.grasso,alberto.valesè,marta.micheli}@unito.it

**Abstract.** In the last decades, the Natural Language Processing (NLP) community has demonstrated committed involvement in addressing societal challenges, particularly in the realm of hate-speech detection. Despite advancements, these phenomena continue to perpetrate, especially online, where users on social network platforms often find themselves in unsafe and possibly harmful environments. Among the various manifestations of hate speech and offensive language, one aspect that has been overlooked by the NLP community is body-shaming. Despite its prevalence among hateful users and its potential to harm a diverse group of individuals, from women to people with disabilities, efforts to counteract this damaging phenomenon remain limited. In this work, we first introduce a novel taxonomy designed to distinguish and classify instances of body-shaming by the targeted group. Following this, we present a dataset of Instagram comments for body-shaming detection and classification in the Italian language, which has been manually annotated according to the taxonomy. After detailing the data-gathering and annotation process, we present a classification benchmark using three BERT-based models to showcase our dataset’s classification potential. Results demonstrate good performances in detecting body-shaming instances across several categories of our proposed taxonomy.

**Keywords:** Natural Language Processing · Body-Shaming · Hate Speech  
**Content Warning:** this paper contains examples of body-shaming, including potentially distressing comments on appearance, gender identity, sexual orientation, ethnicity, and ability. Please proceed with caution.

## 1 Introduction

Body-shaming is the criticism of someone based on the shape, size, or appearance of their body [3]. It can manifest subtly, such as in the form of advice (e.g., medically-based advice from a friend: ‘You should reduce your weight to prevent high blood pressure’) or explicitly, through malevolent insults (e.g., from an unknown social media follower: ‘You need some meat on your bones’) [37].

---

\* Corresponding Author.

The often unmeetable beauty standards imposed by society have a long-lasting tradition, where people have been frequently judged based on their appearance. Moreover, body-shaming often intersects with other forms of discrimination, such as racism [42], misogyny [35], ableism [32], or transphobia [8], and can be seen as one expression of such toxic behavior. With the proliferation of online social platforms, particularly among young users, the phenomenon of body-shaming, like other hate-speech phenomena and cyberbullying, is amplified as social platforms serve as a sounding board for propagating toxic behavior, due to the anonymity of virtual exposure compared to in-person interactions. Body-shaming involves unsolicited, mostly negative opinions or comments about a person’s body [37]. Its manifestations can be multifaceted, targeting aspects like size, shape, weight, body parts, body-related appearance, extremities, etc. Body-shaming has been recognized as a form of misogynistic speech and sexism-related discrimination [26], with women being a primary target due to increased objectification and societal beauty standards.

Many studies have addressed the consequences of body-shaming on people’s health and behavior. For instance, it can lead to low self-esteem, depressive symptoms [18], or disordered eating patterns [15]. Automatic detection of such abusive behavior can be crucial in mitigating and stopping the propagation of this damaging content, eventually resulting in a positive impact on society. So far, the efforts in this direction by the Natural Language Processing (NLP) community have been limited, with only one work in the literature specifically targeting body-shaming detection in English [31]. While the literature presents several works addressing hate speech phenomena in online communities [28], with corpora designed to capture expressions of racism [39], sexism [33], and homophobia [40], this spotlight on specific discriminations highlights a broader issue within the field. Notably, ableism, a form of discrimination that impacts individuals with disabilities, often remains overlooked. Despite a few attempts to include people with disabilities among the targeted groups of hateful comments [20, 24], there is, to the best of our knowledge, a notable absence of datasets that specifically target ableist expressions comprehensively. This encompasses not only hate speech directed towards individuals with disabilities *per se*, but also the employment of ableist language to insult individuals irrespective of their disability status. Using expressions like ‘get treatment’ or ‘you look retarded’ to target non-disabled individuals implies that being associated with disability is intrinsically inferior and tantamount to an insult. This practice makes ableism a universal tool for derogation, reinforcing negative stereotypes and the stigmatization of disability.

In this paper, we address these gaps by first introducing a taxonomy of six body-shaming categories, representing our initial contribution. This taxonomy begins by distinguishing between content that constitutes body-shaming and content that does not. Subsequently, it focuses on the targets of derogatory expressions, identifying six prevalent manifestations of body-shaming on social media: fatphobia, skinny-shaming, misogyny/sexism, ableism, racism, and queerphobia. Although it is not exhaustive or highly detailed, defining this taxonomy

marks a critical step toward more nuanced investigations into body-shaming classification.

Our second, and more comprehensive, contribution proposes a dataset of more than 11k Instagram comments for the detection and classification of body-shaming in Italian<sup>1</sup>. To our knowledge, this represents the first work on body-shaming detection within the Italian context and an initial attempt in general to categorize body-shaming instances, including ableism.

We detail the data collection and annotation processes, as well as an evaluation framework involving three different BERT-based language models. The dataset is designed to serve as a reference for researchers and activists alike, aiming not only to support the fight against this phenomenon within specific communities but also to capture its diverse manifestations. The paper is structured as follows. Section 2 provides an overview of the relevant related work. In section 3, we detail the proposed taxonomy and the dataset creation process. Section 4 outlines the annotation schema, process, and results. In Section 5, we describe classification experiments, while Section 6 concludes the paper.

## 2 Related Works

The NLP community has a strong tradition of focusing on societal challenges, with several efforts made to develop tools for detecting, and consequently preventing, hate speech, cyberbullying, and related phenomena [28, 4]. However, while psychology and social science have already directed attention towards body-shaming [37, 7], in computational linguistics this topic remains largely unaddressed, offering few resources for its detection.

Among the available works, [26] construct a dataset for sexism categorization, including body-shaming as one of the categories of sexism. [21] explore the identification of body-shaming comments through sentiment analysis and classification techniques. [31] provides, to our knowledge, the only available dataset specifically for body-shaming detection, classifying Instagram comments as body-shaming or not. However, the dataset primarily includes ‘indirect’ instances, where comments report on or complain about body-shaming events, rather than being direct expressions of body-shaming. [12] use Naive Bayes Classifier approach to do sentiment analysis on body-shaming tweets in Indonesian. Regarding the realm of harassment and toxicity towards specific discriminated or marginalised groups, many works address racism [23, 39] and misogyny/sexism [33, 26] detection, whereas there are still few works for other targeted hate speech groups. For ableism detection, [20, 24] and [17] investigate the presence of explicit bias against people with disability in sentiment analysis, toxicity models, and LLMs; [25] and [11] include ‘people with special needs’ and ‘handicap’ as target group and category for hate speech, respectively. Among NLP works that address queerphobia<sup>2</sup>, [38] investigate evidence of bias against queer identities in sentiment analysis tools; [9] propose a shared task on homo/transphobia detection

<sup>1</sup> The dataset is available here: <https://github.com/ValeseA/BS-Detect>

<sup>2</sup> Understood as hateful expression and discrimination against LGBTQ + individuals.

in social media; [40] propose a corpus for detecting LGBT+Phobia in Mexican Spanish. Regarding toxicity detection towards specific body shapes, [39] considers fatphobia in their Brazilian Portuguese corpus. Despite extensive literature on discrimination, body-shaming research remains limited, where a dataset for detecting body-shaming in Italian is still lacking, and so is a categorization of body-shaming instances according to the targeted group.

### 3 Dataset Design and Creation

#### 3.1 Body-shaming Taxonomy

Our goal was to create a resource for detecting body-shaming hate speech in a broad sense, targeting a wide range of individuals and diverse expressions of aspect-based critiques. Body-shaming often overlaps with other types of discrimination, particularly targeting individuals from discriminated or marginalized groups. It manifests in various forms, including racism [42], misogyny and sexism [35, 26], ableism [32] and transphobia [8]. Accordingly, we aimed to (i) distinguish instances that constitute body-shaming from those that do not and (ii) classify body-shaming instances based on individuals frequently targeted due to societal standards and widespread discriminatory attitudes.

Therefore, we developed a taxonomy with two hierarchical levels for detecting body-shaming content, drawing inspiration from the Wheel of Power, Privilege, and Marginalization by Sylvia Duckworth<sup>3</sup> [1]. This tool, designed to facilitate discussions about intersectionality and systemic inequality [22], categorizes key social identities and categories, including race, gender, sexual orientation, body size, and ability, to delineate the distribution of societal privilege and marginalization [34], making it a suitable reference for body-shaming. From this literature reference, we created the first draft of the taxonomy, further refined through a grounded theory approach [19] with empirical data from our dataset (see Section 4) to adjust the schema. This taxonomy mirrors the annotation scheme used for our dataset annotation, as illustrated later. It comprises two levels:

***Binary Body-shaming*** The first level establishes a binary categorization between content that constitutes body-shaming versus content that does not. We identify body-shaming as any form of explicit or implicit criticism, humiliation, or derision related to an individual’s physical appearance, persisting regardless of context or intent [37, 3]. This includes negative comments or unsolicited advice about body shape, weight, height, facial features, or any other physical characteristics. It also encompasses derision of ways of speaking, moving, and overall attitude, as these too relate to the body and its presentation.

<sup>3</sup> The original 2020 version is sourced from the author’s Flickr page-<https://www.flickr.com/photos/sylviaduckworth/50500299716/>; a simplified adaptation by the Canadian Council of Refugees (CCR)<https://ccrweb.ca/en/anti-oppression>. Several versions of the wheel have been developed for various contexts.

**Category of Body-shaming** The second level distinguishes six distinct categories of body-shaming that we identified and defined based on the target group: *Fatphobia*; *Skinny-shaming*; *Misogyny/sexism*; *Racism*; *Ableism*; *Queerphobia*. Below we provide definitions and examples for each category. Note that this taxonomy does not claim to be exhaustive of every target of body-shaming, but it serves as a starting point to capture the diverse dynamics of this phenomenon:

- **Fatphobia:** Criticism or negative comments, often delivered as unsolicited advice or health concerns, targeting individuals with bodies that do not conform to societal standards of size and shape, perpetuating the stigma around body diversity and implying that deviation from these standards is undesirable [5] (e.g. ‘You look like a pig’; ‘Someone your size shouldn’t wear that’).
- **Skinny-shaming:** Negative remarks or disparagement directed at individuals perceived as too thin, suggesting they lack health or attractiveness due to their slimness [2] (‘You look sick, eat something!’; ‘Put some meat on your bones’). This category was not included in the initial phase of drafting our taxonomy. However, after encountering several detrimental and hateful comments targeting this specific body shape, we decided to incorporate it into our taxonomy and annotation scheme.
- **Misogyny/sexism:** Body-shaming that specifically targets women, often through critiques that enforce narrow societal beauty standards or demean women for not adhering to these standards, reflecting gender-based prejudice [30] (‘Don’t you see those hairy underarms are gross?’; ‘You’re so flat, real women should have curves’).
- **Racism:** Body-shaming intertwined with racial prejudices, targeting individuals based on racial or ethnic characteristics, including skin color, hair texture, facial features, or body shape typical to specific ethnic groups [6] (‘People like you look like monkeys’; ‘You definitely cannot be Italian with that wig on your head’).
- **Ableism:** Criticism or humiliation based on physical differences, cognitive divergence, or motor abilities, directed at both individuals with disabilities and those without, using negative comparisons to disabilities as a form of insult [32] (e.g. ‘Are you retarded?’; ‘You look like you have Down syndrome’). This practice perpetuates ableist prejudices, diminishing the individual beyond physical and cognitive norms.
- **Queerphobia:** Body-shaming targeting individuals based on their sexual orientation or gender identity, critiquing not only their physical appearance but also their ways of speaking, moving, and overall attitude [8] (e.g. ‘You’re such a horrific freak’; ‘Disgusting, are you even male or female?’). This form of shaming reinforces stereotypes and negates their identity and expression, often scrutinizing these aspects to demean or invalidate the individual’s authentic self.

### 3.2 Data Collection

Our aim was to collect textual data, specifically user comments, from a popular social platform that could be qualified as instances of body-shaming and targeting a vast range of users, including individuals of marginalized groups or those typically targeted for body-critiques. Particularly, we intended to include potentially hateful comments that might fall under—but not limit our focus to—the six categories of body-shaming hate speech groups detailed above. To this aim, similar to [31], we identified Meta’s Instagram<sup>4</sup> as a preferred source for textual data, specifically user comments. This platform, highly popular particularly among teenagers and young adults, is predominantly used for picture posting or “reels”<sup>5</sup>, making it likely for users to expose their personal image and body, and consequently attract hateful comments targeting their physical appearance [14]. Our goal was to gather a wide variety of body-shaming instances, from fat-shaming to transphobic expressions. To direct our search, we thus employed a combined strategy. We selected posts (either pictures or reels) by browsing highly popular open user profiles likely targets of cyberbullying and hate speech, also considering their high exposure. These included: feminist queer pages, advocates of “body positivity”, famous disabled activists, popular non-Caucasian Italian athletes or players, female influencers, public figures with non-conforming bodies, openly transgender activist user pages<sup>6</sup>. Moreover, we combined diverse hashtags to target posts from popular educational, entertainment, satirical, or news Italian accounts whose content could attract hateful messages. The hashtags included: #disabile #sindromedidown #bodypositivity #bodyshaming #modellacurvy #lgbtqitalia #orgoglioqueer. Prior to data crawling, we manually inspected the comment section of the browsed posts and pages to verify the presence of hateful messages. We then carefully selected a total of 100 posts expected to contain such comments, aiming for a balanced distribution across the six identified categories of body-shaming. This aimed to ensure that each subcategory - fatphobia, skinny-shaming, misogyny/sexism, racism, ableism, and queerphobia - was adequately represented in our dataset. To facilitate the data gathering process, we used ExportGram and ExportComments<sup>7</sup>, tools designed for exporting social media comments (including Instagram), to collect potentially abusive content. The data extraction from the 100 selected posts led to a total of 39,467 exported comments.

<sup>4</sup> <https://www.instagram.com/>

<sup>5</sup> Instagram’s “reels” are short videos up to 60 seconds long.

<sup>6</sup> All comments analyzed in this study were extracted from public Instagram profiles, defined as data accessible without the need to log in, and were collected in accordance with Meta’s privacy policy for academic research purposes. Additionally, we have chosen not to disclose specific names from which comments were extracted to uphold privacy and ethical research practices and prevent potential harm. Exceptions may only occur with explicit consent and where necessary for research integrity.

<sup>7</sup> <https://exportcomments.com/>; <https://exportgram.net/>

### 3.3 Data Cleaning and Dataset Creation

After collecting the comments, we carried out basic pre-processing steps to enhance data quality. We started with removing duplicates (identical comments), comments composed solely of emojis, hashtags, tags, gifs, URLs and user mentions (defined by the prefix @). Additionally, we filtered out empty comments and those not in the Italian language<sup>8</sup>. URLs and user mentions were also removed from the remaining comments. Post-cleaning, we obtained 29,003 comments eligible for manual annotation.

Reliably measuring the frequency of abusive content in natural online environments is challenging, with estimates possibly as low as 0.1% to 1% [41]. Recognizing that body-shaming is only a subset of abuse, we chose not to employ sampling techniques used in prior work such as keyword [13] or lexicon-based ones [16]. These methods could risk overlooking nuanced or less overt instances of body-shaming, limiting the diversity and representativeness of our dataset. While aware that this decision might lead to a dataset class imbalance, we prioritized capturing the broadest and most heterogeneous examples of body-shaming. To mitigate potential imbalances as much as possible and ensure a comprehensive overview of body-shaming expressions, we attempted to balance the dataset for annotation by randomly selecting an equal number of comments from each targeted subcategory within the 100 posts. From the initial pool of 29,003 comments, we curated a final sample of 13,212 comments, aiming for a manageable yet diverse set suitable for manual annotation and classification tasks. This approach, while not without its challenges, was intended to minimize bias toward any specific category of body-shaming, attempting a broad representation of expressions within our analysis.

## 4 Dataset Annotation

### 4.1 Annotation Scheme

For the annotation of our dataset, we adhered to a ‘prescriptivist approach’ in data annotation, as we wanted the annotators to refer to our detailed annotation guidelines rather than relying on their subjective interpretations, as far as possible [36]. The annotators were thoroughly instructed with comprehensive guidelines outlining the objectives, specifics of the annotation schema, definitions, clarifications, examples, considerations for borderline cases, and overall instructions. The annotation scheme mirrors the taxonomy illustrated in section 3.1, whose details and categories’ definitions were also presented in the guidelines. Thus, the annotation scheme we developed was composed of two levels:

---

<sup>8</sup> Comments in languages commonly spoken in Italy but not recognized as dialects of Italian, such as Neapolitan or Sicilian, were excluded. Comments that included so-called ‘regional Italian’ or dialects of Italian, such as Tuscany and Roman expressions, were instead included.

- *Body-shaming task*: In this first binary level, each comment was labeled as either containing an instance of body-shaming or not, using the labels **yes** and **no**. The annotation was prompted by the question, “Does the comment include a body-shaming instance?”. If labeled as **yes**, a further multilabel classification could follow, when applicable. Importantly, annotators were instructed to exclude generic hate speech not specific to body-shaming. Only comments that could reasonably be interpreted as body-shaming had to be categorized as such.
- *Categorization task*: At the secondary multilabel level, comments identified as body-shaming could be further tagged with specific labels representing one of the six types of discrimination outlined in our taxonomy. The available labels at this level were: **fatphobia**, **skinny-shaming**, **misogyny/sexism**, **racism**, **ableism**, and **queerphobia**. If none of these categories were applicable, the annotation at this level could be left blank. Moreover, the annotators were allowed to select up to two labels. This scenario arose when the comment fell into one of the following situations: (i) the content could be relevant to either one of two categories, but it was unclear which category it fit more accurately, or (ii) the content clearly pertained to both categories of discrimination simultaneously. This occurred, for example, in cases where a woman’s body was derogatorily commented as both being too skinny and not adhering to traditional female beauty standards, e.g. ‘Where’s your chest, skeleton?’(this intersects misogyny/sexism and skinny-shaming categories). Another instance is when a non-cisgender individual with a non-conforming body received comments such as ‘Aren’t you ashamed of yourself for being such a fat faggot?’, exemplifying the overlap of fatphobia and queerphobia.

Comment	Body-sh.	Category
“Bellissimo il costume! Che marca è?” (Beautiful swimsuit! What brand is it?)	no	-
“Che fisico di merda si può dire?” (What a shitty physique can one have?)	yes	-
“Con la 4’ di seno staresti meglio...” (With a size D breast, you would look better...)	yes	misogyny/sexism
“6 tozza ed hai la faccia da down” (You’re stocky and have a Down syndrome face)	yes	ableism misogyny/sexism
“Muori di obesità e HIV” (Die of obesity and HIV)	yes	fatphobia queerphobia
“Come i tratti somatici di muso di cavallo” (As the somatic features of a horse’s muzzle)	yes	racism

Table 1: Examples of dataset comments with their annotation.

In Table 1, we report some examples of comments with their expected label, according to our developed annotation scheme.



## 4.2 Annotation Process

The annotation was carried out by three expert annotators, all of whom are Italian native speakers with prior experience and expertise in hate-speech annotation tasks. One annotator is also co-author of this paper. After being provided with the guidelines, a discussion session was held to address any questions or clarifications regarding the annotation criteria. The guidelines also specified that annotators could skip comments that did not meet the criteria required for the annotation, such as comments containing only emojis or non-interpretable text (e.g., “ahahah”), or comments not in Italian. These comments would be subsequently excluded from the dataset after completion of the task. In the annotation process, annotators were provided with the sources of the Instagram comments to assist in accurately contextualizing each comment. This step was crucial for understanding the nuances and intent behind the remarks and distinguishing between comments that are explicitly or implicitly body-shaming and those that are not. To ensure alignment among the annotators and a shared understanding of the guidelines, a pilot annotation of 25 randomly selected comments was conducted. This preliminary task allowed annotators to discuss any discrepancies and refine the guidelines if necessary. After aligning on the annotation task during the pilot phase, the annotators performed the main annotation task on the dataset. Throughout the process, we maintained close collaboration with our annotators, ensuring their feedback was integrated into our guidelines and their well-being was consistently monitored and safeguarded.

## 4.3 Annotation Results

After completing the annotation, the dataset was cleansed of any comments that were skipped, resulting in a total of 11,393 annotated comments for the final dataset.

Task	Body-Sh.	Fatph.	Skinny-Sh.	Misog./Sexism	Racism	Ableism	Queerph.
Fleiss’ $\kappa$	0.694	0.611	0.628	0.182	0.721	0.677	0.567

Table 2: IAA Scores for the Body-Shaming and Categorization Tasks. The first score represents consensus on the presence of body-shaming, while subsequent scores indicate agreement levels for specific categories of body-shaming, calculated with an adapted binary Fleiss’  $\kappa$  for multilabel tasks.

**IAA Measurement** We measured the Inter-Annotator Agreement (IAA) among the annotators using Fleiss’ Kappa. Below, we briefly discuss the IAA values obtained for our dataset, presented in Table 2.

**Body-Shaming task:** For the first binary task we achieved a Fleiss’  $\kappa$  score of 0.694. This score indicates substantial agreement among annotators, suggesting that the annotators were generally consistent in identifying whether a comment contained body-shaming content, also indicating clear task guidelines. However,

this score also hints at the inherent subjectivity involved in identifying body-shaming instances within some comments. We found that this subjectivity was especially pronounced in cases where body-shaming was implicit rather than explicit (e.g. “Guarda che l’obesità è una malattia!” *Just so you know, obesity is a disease!*; “Le belle ragazze sono altre” *The pretty girls are others*), or when comments resided on the borderline between body-shaming and generic insults (e.g. “Ma non ti fai schifo da solo?” *Don’t you disgust yourself?*; “Fenomeno da baraccone” *You’re a freak*). Subjectivity arose also as linked to individual differences in sensitivity to certain types of remarks. For instance, unsolicited comments framed as compliments (e.g. “Quelle smagliature ti donano!” *Those stretch marks suit you!*), sarcastic comments, or observations that might imply a negative remark (“Sembri incinta” *You look pregnant*) were interpreted variably. Some annotators saw these as covert forms of body-shaming, while others considered them as innocuous or genuinely positive.

**Categorization task:** For the secondary level’s multilabel task, where comments could be tagged with up to two different labels, we adapted the Fleiss’  $\kappa$  calculation to account for multiple labels per comment. This involved transforming our dataset into a binary decision matrix for each label, allowing us to systematically account for instances where annotators agreed on at least one of the potential labels. Each of the six labels was considered a separate decision, marked as present (1) or absent (0) for each comment by each annotator. This approach allowed us to capture partial agreements among annotators, especially relevant for comments that spanned multiple discrimination categories. This binary decision matrix enabled us to compute the IAA in a manner that captures both full and partial agreement among annotators. The calculated IAA scores indicate substantial agreement among annotators for almost all categories, ranging between 0.567 and 0.721. However, notable disparity is observed in the category of misogyny/sexism, which exhibits significantly lower agreement. This disparity will be further discussed in the subsequent paragraph.

Label	Annotator 1	Annotator 2	Annotator 3
yes	1268	1154	1381
no	10125	10239	10012
fatphobia	537	505	639
skinny-shaming	55	51	78
misogyny-sexism	84	59	419
racism	30	19	30
ableism	136	143	150
queerphobia	151	260	173

Table 3: Label distribution for Body-Shaming detection and Categorization by annotator: the numbers refer to counts of comments identified under each label.

**Label Distribution** Table 3 reports the label distribution across the two annotation levels for all three annotators.

The *Body-shaming* task reveals a significant class imbalance, with the majority of comments labeled as **no** for lacking body-shaming content. This distribution aligns with the expected prevalence of neutral comments over explicit hateful instances, given the context of social media. The observed imbalance was anticipated and is considered acceptable for our study’s goal, which was developing a nuanced understanding for body-shaming content, rather than achieving a perfectly balanced dataset. Despite the class imbalance, the quantity of **yes** labels still provides a robust foundation for the analysis of body-shaming content, supporting the validity of our subsequent analyses and model training. The consistent number of **yes** labels across annotators indicates a good agreement on body shaming, enhancing dataset reliability.

In the *Categorization Task*, label distribution reveals insights into both the prevalence of specific types of body-shaming within the dataset and the consensus among annotators. **fatphobia** is consistent among annotators, indicating its prevalence in the dataset and its clear definition in the guidelines. **skinny-shaming** and **racism** are less common, with Annotator 3 showing slight variance in identifying the former. **misogyny-sexism** exhibits considerable variance, notably with Annotator 3 identifying significantly more instances. This may stem from the challenge in differentiating between misogyny/sexism-related and “general” body-shaming when the targets are women. This discrepancy also explains the low IAA for this category shown in Table 2. Finally, **ableism** and **queerphobia** are more frequent, with good distribution among annotators (despite Annotator 2’s slight predisposition for this label), indicating both a significant presence in the dataset and a clear guideline comprehension.

## 5 Evaluation

To assess our dataset’s reliability, we conducted a comprehensive evaluation, framing the detection of body-shaming and its categorization, when present, as binary classification tasks. These experiments aimed to (i) gauge how well state-of-the-art language models identify body-shaming instances and (ii) set benchmarks for future research, marking the first effort to classify body-shaming in Italian. Utilizing a dataset derived from annotations and a majority voting system from three annotators, data was included in the training set only when at least two annotators agreed.

Given the dataset’s imbalance towards non-body-shaming instances, as shown in Table 3, we selected a balanced subset of 3,000 comments to ensure a 60-40 split between non-body-shaming and body-shaming instances. Comments were standardized to 32 tokens in length to align with the average comment size of 14 tokens for more effective model training.

For the second-level categorization, we treated it as six separate binary classifications to avoid biases from the uneven label distribution, which could skew the model’s learning focus. Thus, only comments explicitly marked as involving body-shaming were used for this task, concentrating the training on identifying specific categories of body-shaming comments.

We fine-tuned three pre-trained BERT-based models from the Hugging Face platform<sup>9</sup> for our tasks: UmBERTo[27], a widely-used Italian BERT model; AIBERTo[29], pre-trained on Twitter data to potentially enhance performance on Instagram; and XLM-RoBERTa[10], a Multilingual Language Model, to evaluate its adaptability. The fine-tuning process involved training for a maximum of 30 epochs with warmup steps set at 20% and implemented early stopping to prevent overfitting. The batch size was limited to 16 for model convergence and memory constraints, and a learning rate of 5e-5 was selected based on literature recommendations for similar models.

The fine-tuning results, including precision, recall, accuracy, and F1-score metrics are detailed in Table 4. These metrics demonstrate a strong capability in body-shaming recognition, with all models performing well. Particularly noteworthy is AIBERTo’s superior performance, possibly attributed to its Twitter-based pre-training, which aligns well with the social media context of our dataset. For the second task, it is observed that the categories of skinny-shaming and misogyny/sexism lack data, as none of the models achieved performance better than random chance. However, for other categories, the best-performing model and its corresponding metric values are presented. In particular, UmBERTo excels in ableism and fatphobia detection, while AIBERTo performs best in racism and queerphobia identification. XLM-RoBERTa generally lags behind, possibly due to its multilingual training not being focused on Italian. Conversely, AIBERTo consistent performance may be attributed to its Twitter data training, aligning closely with the language used on Instagram. Notably, categories with more data instances tend to yield better results. Yet, the model’s proficiency in identifying racism may reflect the distinct language patterns specific to body-shaming within this category. However, the limited number of instances in this category warns of a potential for overfitting, despite impressive metrics.

Task	Model	Precision	Recall	F-1 score	Accuracy
Body-shaming	AIBERTo	<b>0.81</b>	<b>0.81</b>	<b>0.81</b>	<b>0.81</b>
	UmBERTo	0.80	0.80	0.80	0.80
	XLM-RoBERTa	0.80	0.80	0.80	0.80
Fatphobia	AIBERTo	<b>0.75</b>	0.73	0.74	<b>0.76</b>
	UmBERTo	0.74	<b>0.75</b>	<b>0.75</b>	0.75
	XLM-RoBERTa	0.72	0.72	0.72	0.73
Racism	AIBERTo	1.00	0.75	0.83	1.00
Ableism	AIBERTo	0.78	<b>0.78</b>	0.78	0.92
	UmBERTo	<b>0.91</b>	0.76	<b>0.81</b>	<b>0.94</b>
Queerphobia	AIBERTo	0.75	<b>0.71</b>	<b>0.73</b>	<b>0.88</b>
	UmBERTo	<b>0.84</b>	0.61	0.64	<b>0.88</b>

Table 4: Results of the classification tasks for models with statistically significant performance. Performance metrics for tasks where models did not achieve better than random chance are omitted.

<sup>9</sup> <https://huggingface.co/>

## 6 Conclusion

In this work, we presented a novel two-level taxonomy for detecting and classifying body-shaming content, covering six distinct categories: Fatphobia, Skinny-shaming, Misogyny/sexism, Racism, Ableism, and Queerphobia. Our main contribution is the first dataset for body-shaming detection and classification in Italian, featuring 11,393 Instagram comments annotated according to our detailed taxonomy. Notably, our focus on Ableism detection introduces a new dimension to body-shaming research. Classification experiments with three BERT-based models, including two Italian-specific and one multilingual, yielded encouraging results, highlighting the efficacy of language-specific models, especially social-media-adapted, for accurately identifying body-shaming instances. This taxonomy and the dataset aim to advance the understanding and mitigation of body-shaming, serving as resources for both researchers and activists.<sup>10</sup>

## References

1. Andersen, N.: Diverse examples and balanced perspectives. Enhancing Inclusion, Diversity, Equity and Accessibility (IDEA) in Open Educational Resources (OER) (2022)
2. Anderson, J., Bresnahan, M.: Communicating stigma about body size. *Health Communication* **28**, 603 – 615 (2013)
3. Arumugam, N., Manap, M.R., Mello, G.D., Dharinee, S.: Body shaming: Ramifications on an individual. *International Journal of Academic Research in Business and Social Sciences* (2022)
4. Bassignana, E., Basile, V., Patti, V., et al.: Hurtlex: A multilingual lexicon of words to hurt. In: CEUR Workshop proceedings. vol. 2253, pp. 1–6. CEUR-WS (2018)
5. Brewis, A., Wutich, A., Falletta-Cowden, A., Rodriguez-Soto, I.: Body norms and fat stigma in global perspective. *Current Anthropology* **52**, 269 – 276 (2011)
6. Capodilupo, C.M., Kim, S.: Gender and race matter: the importance of considering intersections in black women’s body image. *Journal of counseling psychology* **61** **1**, 37–49 (2014)
7. Cassidy, L.: Body shaming in the era of social media. *Interdisciplinary Perspectives on Shame: Methods, Theories, Norms, Cultures, and Politics* **157**, 396 (2019)
8. Castellini, G., Rossi, E., Cassioli, E., Sanfilippo, G., Ristori, J., Vignozzi, L., Maggi, M., Ricca, V., Fisher, A.D.: Internalized transphobia predicts worse longitudinal trend of body uneasiness in transgender persons treated with gender affirming hormone therapy: a 1-year follow-up study. *The Journal of Sexual Medicine* **20**(3), 388–397 (2023)
9. Chakravarthi, B.R., Kumaresan, P.K., Priyadharshini, R., Buitelaar, P., Hegde, A., Shashirekha, H.L., Rajiakodi, S., García-Cumbreras, M., Jiménez-Zafra, S.M., García-Díaz, J.A., Valencia-García, R., Ponnusamy, K.K., Shetty, P., García-Baena, D.: Overview of third shared task on homophobia and transphobia detection in social media comments. In: Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion. European Chapter of the Association for Computational Linguistics, Malta (March 2024)

<sup>10</sup> We would like to express our gratitude to Sowelu for his crucial contribution and to Roger for his wise help.

10. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
11. Del Vigna<sup>12</sup>, F., Cimino<sup>23</sup>, A., Dell’Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on facebook. In: Proceedings of the first Italian conference on cybersecurity (ITASEC17). pp. 86–95 (2017)
12. Diantoro, K., Sitorus, A.T., Rohman, A., et al.: Analyzing the impact of body shaming on twitter: A study using naive bayes classifier and machine learning. *Digitus: Journal of Computer Science Applications* **1**(1), 11–25 (2023)
13. Elsherief, M., Belding-Royer, E.M., Nguyen, D.: #notokay: Understanding gender-based violence in social media. In: International Conference on Web and Social Media (2017)
14. Fitria, K., Febrianti, Y.: The interpretation and attitude of body shaming behavior on social media (a digital ethnography study on instagram) **3**, 12–25 (2020)
15. Flak, S.R.: The influence of maternal body-shaming comments and bodily shame on portion size. Ph.D. thesis, University of South Florida (2021)
16. Frey, T.F., Fernández, M., Novotný, J., Alani, H.: Exploring misogyny across the manosphere in reddit. Proceedings of the 10th ACM Conference on Web Science (2019)
17. Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, E., Brewer, R.: “i wouldn’t say offensive but...”: Disability-centered perspectives on large language models. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. p. 205–216. FAccT ’23, Association for Computing Machinery, New York, NY, USA (2023)
18. Gam, R.T., Singh, S.K., Manar, M., Kar, S.K., Gupta, A.: Body shaming among school-going adolescents: prevalence and predictors. *International Journal of Community Medicine and Public Health* **7**, 1324–1328 (2020)
19. Glaser, B., Strauss, A.: Discovery of grounded theory: Strategies for qualitative research. Routledge (2017)
20. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., Denuyl, S.: Social biases in nlp models as barriers for persons with disabilities. arXiv preprint arXiv:2005.00813 (2020)
21. Jaman, J.H., Hannie, H., Simatupang, M.R.A.: Sentiment analysis of the body-shaming beauty vlog comments (2020)
22. Kellam, N., Svihla, V., Davis, S.C., Sajadi, S., Desiderio, J.: Using power, privilege, and intersectionality to understand, disrupt, and dismantle oppressive structures within academia: A design case. In: CoNECD Conference (2021)
23. L, S., J, A., E, A.S., M, S.R., N., H.K.: Racism detection using deep learning techniques. E3S Web of Conferences (2023)
24. Narayanan Venkit, P., Srinath, M., Wilson, S.: Automated ableism: An exploration of explicit disability biases in sentiment and toxicity analysis models. In: Ovalle, Anaelia, e.a. (ed.) Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023). pp. 26–34. Association for Computational Linguistics, Toronto, Canada (Jul 2023)
25. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., Yeung, D.Y.: Multilingual and multi-aspect hate speech analysis. arXiv preprint arXiv:1908.11049 (2019)
26. Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., Varma, V.: Multi-label categorization of accounts of sexism using a neural framework. In: Conference on Empirical Methods in Natural Language Processing (2019)
27. Parisi, L., Francia, S., Magnani, P.: Umberto: an italian language model trained with whole word masking (2020)

28. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* **55**, 477–523 (2021)
29. Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., Basile, V.: ALBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. vol. 2481. CEUR (2019)
30. Ramati-Ziber, L., Shnabel, N., Glick, P.: The beauty myth: Prescriptive beauty norms for women reflect hierarchy-enhancing motivations leading to discriminatory employment practices. *Journal of personality and social psychology* (2020)
31. Reddy, V., Abburi, H., Chhaya, N., Mitrovska, T., Varma, V.: 'you are big, s/he is small' detecting body shaming in online user content. In: *Social Informatics* (2022)
32. Reel, J.J., Bucciare, R.A.: Ableism and body image: Conceptualizing how individuals are marginalized. *Women in Sport and Physical Activity Journal* **19**(1), 91–97 (2010)
33. Richter, A., Sheppard, B., Cohen, A., Smith, E., Kneese, T., Pelletier, C., Baldini, I., Dong, Y.: Subtle misogyny detection and mitigation: An expert-annotated dataset. In: *Socially Responsible Language Modelling Research* (2023)
34. Riitaoja, A.L., Virtanen, A., Reiman, N., Lehtonen, T., Yli-Jokipii, M., Udd, T., Peniche-Ferreira, L.: Migrants at the university doorstep: How we unfairly deny access and what we could (should) do now. *Apples - Journal of Applied Language Studies* (09 2022)
35. Roodt, K.: (Re) constructing body shaming: Popular media representations of female identities as discursive identity construction. Ph.D. thesis, Stellenbosch: Stellenbosch University (2015)
36. Röttger, P., Nozza, D., Bianchi, F., Hovy, D.: Data-efficient strategies for expanding hate speech detection into under-resourced languages. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. pp. 5674–5691. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022)
37. Schlüter, C., Kraag, G., Schmidt, J.: Body shaming: an exploratory study on its definition and classification. *International journal of bullying prevention* (2021)
38. Ungless, E.L., Ross, B., Belle, V.: Potential pitfalls with automatic sentiment analysis: The example of queerphobic bias. *Social science computer review* **41**(6), 2211–2229 (2023)
39. Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., Benevenuto, F.: HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: Calzolari, Nicoletta, e.a. (ed.) *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. pp. 7174–7183. European Language Resources Association, Marseille, France (Jun 2022)
40. Vázquez, J., Andersen, S., Bel-Enguix, G., Gómez-Adorno, H., Ojeda-Trueba, S.L.: Homo-mex: A mexican spanish annotated corpus for lgbt+ phobia detection on twitter. In: *The 7th Workshop on Online Abuse and Harms (WOAH)*. pp. 202–214 (2023)
41. Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S.A., Margetts, H.Z.: Challenges and frontiers in abusive content detection. *Proceedings of the Third Workshop on Abusive Language Online* (2019)
42. Williams, S.: The problematic body-shaming of black female athletes in professional sports (2018)