GSI:detect - A Perspectivist Approach to Gender Stereotypes Identification in Italian

Davide Testa[1,2], Sofia Brenna[1,3], Manuela Speranza[1], Gloria Comandini[4], Stefania Cavagnoli[5], Bernardo Magnini[1]

[1]Fondazione Bruno Kessler (FBK), [2]University of Rome La Sapienza, [3]Free University of Bozen-Bolzano
[4]Istituto Italiano di Studi Germanici (IISG), [5]University of Trento
{dtesta, sbrenna, manspera, magnini}@fbk.eu
comandini@studigermanici.it, stefania.cavagnoli@unitn.it

Abstract

The deconstruction of gender stereotypes is essential to prevent discrimination, marginalization and gender-based violence. Despite the increasing attention to this issue, research in this field often focuses on explicitly sexist or hateful communication, leaving out all the cases where stereotypes are produced unconsciously or even with apparently positive intentions. Moreover, the identification and analysis of gender stereotypes is often a very subjective task, heavily influenced by the researcher's background, beliefs and personal sensitivity. In this context *GSI:detect*, a dataset for gender stereotypes identification in Italian, has been annotated following a perspectivist approach that gives value to the different points of view of four annotators. It has been designed to address (i) the lack of resources focusing on naturally occurring and non-hateful language conveying implicit or ambiguous forms of gender stereotypes, and (ii) the scarcity of datasets that can capture multiple interpretations as well as the inherent variation and disagreement in human perception. Baseline experiments with several LLMs confirm the challenging nature and value of such a linguistic resource, revealing both apparent differences and limitations in performance among the evaluated models, and raising questions about the extent to which current LLMs are suitable for detection and classification tasks in this field.

Content warning: Examples taken from the GSI:detect dataset may contain sensitive or potentially distressing content.

Keywords: gender stereotypes, learning from disagreement, dataset, linguistic resources, perspectivism

1. Introduction

In the context of Italian language and culture, stereotypes are studied by many areas of expertise, such as psychology, sociology and, more importantly for this work, linguistics (Arcuri and Cadinu, 1998; Peruzzi et al., 2019b; Cavagnoli and Dragotto, 2021). We can define stereotypes as preconstituted, generalised and simplistic opinions about people, events, and situations. Stereotypical opinions are not based on personal evaluation of individual cases, but are mechanically repeated, creating what has been described as a "culturally constructed cage" (Biemmi, 2020). Therefore, stereotypes hinder critical thinking (Biemmi, 2020) and elicit the insurgence of prejudice, as well as discrimination (Jackson, 2011).

Gender stereotypes (GSs) are stereotypes based on socially constructed beliefs regarding the "appropriate" roles, behaviours, and appearances that people should have according to their assigned gender. Therefore, GSs are a cultural conformist force that pushes individuals to shape their personalities based on specific social expectations, which in the Western world are polarized into two groups: men and women (Peruzzi et al., 2019a)[1]. From this perspective, the two poles are in a hierarchical relationship, with the female pole subordinate to the male pole (Biemmi, 2020). Moreover, deviation from the stereotypical perceived "normalcy" of femininity and masculinity, including the "contaminations" between the two genders (i.e. women who adopt features culturally perceived as masculine: playing soccer, having short hair, not wearing makeup, etc. - and the other way round for men), may result in social marginalization and stigmatization (Peruzzi et al., 2019a).

This situation makes the deconstruction of GSs necessary to prevent discrimination and gender-based violence. Consequently, in recent years the rise of Natural Language Processing (NLP) has made it possible to automatically identify and analyse the linguistic expressions through which GSs are conveyed. However, despite the increasing attention to this issue, research in this field often focuses on explicitly sexist or hateful communication, leaving out all the ambiguous cases where stereotypes are produced unconsciously or even with apparently positive intentions. Moreover, the subjectivity involved in identifying such phenomena has often been simplified through binary annotation

[1]Although our present study focuses on stereotypes targeting women and men, this choice is purely methodological and does not assume in any way a binary view of gender. The analysis of GSs applied to non-binary people might be the object of future works.

Proceedings of NLPerspectives @ LREC 2026, pages 21–32
12 May 2026.

schemes, obscuring the diversity of human point of views and calling for evaluation frameworks that explicitly account for annotator variability.

In this context, we introduce the GSI:detect dataset, a new Italian linguistic resource for the study of GSs in short texts. The dataset is designed to address two main gaps in current research: (i) the lack of resources focusing on naturally occurring, non-hateful language containing implicit or ambiguous forms of GSs, while also contributing to a more comprehensive understanding of the phenomenon by including stereotypes directed at both women and men; and (ii) the scarcity of datasets that adopt a perspectivist approach to annotation, useful for capturing multiple interpretations as well as the inherent diversity and disagreement in human perception of stereotypes. With this purpose in mind, GSI:detect offers a novel resource for both linguistic and computational research. From a linguistic perspective, it allows the exploration of heterogeneous ways in which stereotypes manifest in everyday Italian discourse. From a computational point of view, it provides a challenging test-bed to assess how well Large Language Models (LLMs) are able to identify, interpret, classify, and reason about GSs in Italian, especially in situations where subjectivity plays a central role.

The paper is organized as follows. Section 2 presents some related work, focusing both on GSs in NLP and on the perspectivist approach. In Section 3, we describe the GSI:detect dataset, outlining the data collection procedure, the manual annotation process, and related statistics. Finally, Section 4 details the experimental setup, including task design and evaluation metrics, and Section 5 presents and discusses the results, illustrating the effectiveness of GSI:detect for assessing model performance on these tasks.

2. Related Work

2.1. Gender Stereotypes in NLP

GSs have been studied extensively in NLP; however, for the purposes of this work, we focus on two primary research directions.

The first one is the automatic recognition of GSs in texts produced by humans and, more recently, also produced by LLMs. These investigations are generally applied in the context of hateful sexist communication, and therefore are often associated with hate speech (HS) detection. This is due to the fact that most cases of misogynistic hate speech are also imbued with GSs (Fersini et al., 2018). For example, in Kirk et al. (2023) GSs are a subcategory of sexist animosity, while in Plaza et al. (2023) they are one of the many categories of sexism. Similarly, in the context of the Italian language, stereotypes detection has been mostly a subject in automatic recognition of misogynistic hate speech (Fersini et al., 2018).

The second field is the investigation of GSs encoded in LLMs, due to the presence of prejudiced material in their training datasets. In fact, it has been widely assessed that LLMs are prone to spread stereotypes, prejudices and, more generally, social biases regarding, for example, race, ethnicity, religion, age, sexual orientation and, of course, gender identity (Cao et al., 2022; Ovalle et al., 2023), due to the presence of these biases in the human-produced texts used to train them (Talat et al., 2022).

Taking into account the results obtained from these two fields of research, our investigation aims to go one step further: to discover how well LLMs - given their inclination to spread social biases - perform in recognizing GSs. A similar study has been recently conducted by Mitchell et al. (2025), who focused on the recognition of multilingual social stereotypes by LLMs. The dataset used by Mitchell et al. (2025) is, however, quite different from ours, as it consists of a set of short stereotypical texts [2], paired with non-stereotypical counterparts generated through a template-based process. On the other hand, as will be better explained in Section 3.1, our dataset (in addition to focusing solely on GSs) consists of texts that were not produced specifically for our experiment, but rather texts that were naturally written in uncontrolled online environments. In this way, we wanted to collect not only the gender stereotypes that a group of experts might imagine, but also and above all the unexpected and ambiguous cases that arise in spontaneous writing on the web.

Furthermore, we did not want to investigate GSs only in hate speech. In fact, while they can be found in misogynistic hate speech (Kirk et al., 2023; Fersini et al., 2018), they can appear in non-hateful communication as well, as unconscious stereotypes can also be used with well-meaning intentions by both men and women (e.g., *a woman's intuition is never wrong!*). From this perspective, GSs can be intertwined with the phenomenon of micro-aggressions (Sue, 2010), whose nature of implicit indignity makes them particularly prone to being produced even by their own victims (e.g., from: *I didn't do well [in a science exam], but, oh well, girls aren't supposed to be good at science anyway, ha-ha.* (Harrison and Tanner, 2018).

[2]The texts from Mitchell et al. (2025) were produced by a group of data creators who spoke different native languages and regarding not only GSs, but also prejudices about age, race, physical appearance etc.

2.2. The Perspectivist Approach

Recognizing stereotypes is often a very subjective task, as seen in Sanguinetti et al. (2018) where the inter-annotator agreement between expert annotators for racist stereotypes is 0.41 (Cohen's *k*). In fact, due to their pervasiveness and sometimes implicit nature, highly subjective tasks performed by humans (such as GS or HS recognition[3].) can be heavily influenced by interiorized biases (Basile et al., 2023; Muscato et al., 2024), level of expertise in the task, affinity or belonging to the group victim of prejudice (Wojatzki et al., 2018), and even more generic personal opinions (Klenner et al., 2020).

Therefore, disagreement in annotation is not always caused by lack of attention or other kinds of genuine mistakes, but might be a useful cue for exploring the complexity of human experience with respect to subjective themes. This methodology has been defined a Perspectivist Approach (Basile, 2020; Basile et al., 2023; Rizos and Schuller, 2020; Muscato et al., 2024), as it revolves around the idea that NLP should not rely on the classic gold-standard corpora with a majority-aggregated annotation, but should adopt new strategies in order to integrate the added value of opinion's diversity in annotation. In fact, several studies, see for example Klenner et al. (2020), underline that majority voting may suppress rightful points of view that add new information about a topic, and that they should not be classified as errors just because they are a minority vote.

Given that GS annotation can be a very subjective task, we decided to adopt a Perspectivist approach in the creation of our dataset, aiming to take advantage of the inherent subjectivity of the task to better understand the reasons behind different annotations and to study the ambiguous gray area in the continuum between explicit GSs and non-stereotyped communication. While prior work models annotator disagreement as a full probability distribution (Madeddu et al., 2023), more recent approaches estimate or elicit such distributions directly from models, either via token-level probabilities over a closed label space (Santurkar et al., 2023) or by prompting models to output probability distributions (Pavlovic and Poesio, 2024). In contrast, we adopt a simpler scalar formulation based on the mean of binary judgments across annotators, which preserves disagreement in a compact form while ensuring a model-agnostic evaluation setting, particularly for closed decoder-based models where access to probability distributions is limited.

[3]However, it is important to underline that there are also different opinions on the subjectivity of HS, such as (Cercas Curry et al., 2024). Moreover, while there is a level of subjectivity in the annotation of phenomena such as GSs and HS, this subjectivity does not deny the harm done by HS and GSs, nor the fact that there are certain kinds of hateful or stereotyped texts that are recognised as such with an extremely high agreement. This is the case of, for example, extremely sexist HS (e.g. *women like to be raped*), which is recognized as such by both male and female annotators, despite their disagreement on more subtle forms of sexism (e.g. *female quotas are useless*) (Wojatzki et al., 2018).

3. The GSI:detect Dataset

GSI:detect is a new Italian resource developed for the detection, analysis and classification of gender stereotypes in written texts; it has been used as the reference dataset for a shared task with the same name organized within the Evalita 2026 Evaluation Campaign[4] (Comandini et al., 2026) and is distributed under a *Creative Commons NonCommmercial-ShareAlike License*. The dataset is publicly available on its official GitHub repository.

It consists of 1,010 short written Italian texts (for a total of 52,118 tokens), collected to capture authentic, naturally occurring language and to represent a wide range of communicative contexts in which gender stereotypes may appear at different levels of prototypicality, or not at all.

3.1. Data Collection

The texts included in GSI:detect have been manually collected from both social media and informative websites, in order to provide a balanced representation of both formal and informal written Italian.

We collected comments from discussion threads or articles related to different topics from the following social media:

- social media pages discussing gender issues from diverse ideological perspectives[5];
- Facebook and Instagram pages of major Italian newspapers[6] and sports newspapers (*La Gazzetta dello Sport* and *Eurosport Italia*);
- public Facebook groups related to chess and mathematics;

[4]https://gsi-d-evalita.fbk.eu/

[5]Such as the Instagram page of the feminist influencer Chiara Becchi Manzi, the pick-up artists agency *Playlover Academy* and the "mom influencer" *amoree__di_mamma*, as well as from the Facebook pages of ironic and parodic groups such as *Alpha Woman* and *La società femminista*.

[6]*Domani* and *Il Post* from Instagram, and *ANSA*, *Il Corriere della Sera*, *La Repubblica*, *La Stampa*, *La Verità*, *Open* and *SkyTG24*.

Labels	GS value	Example
no-no-no-no	0	Non comprendo come si possano paragonare due fenomeni, gravissimi entrambi e concordo, come femminicidi e morti sul lavoro. *(I don't understand how one can compare two phenomena, both very serious, and I agree on that, such as femicides and workplace deaths.)*
yes-no-no-no	0.25	Tenete duro ancora qualche giorno e i vostri fidanzati partiranno in vacanza con le loro mogli. *(Hold on for a few more days and your boyfriends will be going on vacation with their wives.)*
yes-yes-no-no	0.50	Io rimango dell'idea che un figlio ha sempre bisogno della sua mamma, anche per dire buongiorno e buona notte. E la mamma idem. Soprattutto la mamma *(I still think an [adult] child always needs his/her mother, even to say good morning and good night. And the mother too. Especially the mother*
yes-yes-yes-no	0.75	[Commento ad articolo di giornale dal titolo "Negli Usa quasi un manager su due è donna. In Italia meno di 1 su 3"] Infatti il Made usa va' peggio del Made italy *([Comment on a newspaper article titled "In the US, almost one in two managers is a woman. In Italy, less than one in three"] In fact, Made in USA is doing worse than Made in Italy)*
yes-yes-yes-yes	1	[Rivolto a una utente donna] fatevi voi una doccia e copritevi. Le donne vere si coprono. Gli animali vanno in giro nudi. *([Addressed to a female user] Take a shower and cover up. Real women cover up. Animals go around naked.)*

Table 1: GS values corresponding to the five possible combinations of labels assigned by the four annotators (non aggregated labels).

- more generic sources on Facebook such as gossip pages (*Cosa?*) and pseudo-scientific speculations (*Ghiandola pineale - Il terzo occhio*)
- Reddit pages focusing on dating and relationships (such as *dating_advice*).

As far as informative websites are concerned, we collected excerpts from minor websites spanning from women-oriented spaces (e.g. *Femal*) to local Christian websites (e.g. *Amici del Timone*).

This variety allows us to explore GSs in different contexts (e.g., family, sport, politics, etc), as presented in detail in Table 2.

The resulting dataset includes texts that may or may not display stereotypical content within a linguistic structure that distinguishes between two types of items:

- NO CONTEXT texts, which can be understood without additional contextual information (Examples 1 and 3);
- WITH CONTEXT texts, which are not self contained and are therefore enriched with standardized human-generated metadata, which usually contains contextual information such as the headline of a newspaper article (Example 2).

Furthermore, the dataset includes not only GSs regarding women (Example 1), but also GSs regarding men (Example 2), or both (Example 3).

1. Vabbè oggettivamente le femmine su alcune cose non sono in grado. XD Fagli cambiare una ruota di scorta XD

   (Come on, females are objectively incapable of some things. XD Have them change a tire XD)

2. (Commento ad articolo di giornale dal titolo "Il corpo di ballo di Marco Mengoni balla al ristorante sulle note di 'Mi fiderò'") Li vedo bene in guerra contro i russi. XDXDXD

   ([Comment to a newspaper article titled "Marco Mengoni's dance troupe dances at the restaurant to the notes of 'Mi fiderò'"] I can see them doing well at war against the Russians XDXDXD)

3. Paga l'uomo... Se paga lei è lei l'uomo della coppia.

   (The man should pay... If she pays, she is the man of the couple.)

Topic	N. of texts	Percentage
Family	95	9%
Gender	167	17%
Gossip	115	11%
Politics	88	9%
Romance	68	7%
Sport	138	14%
Violence	67	7%
Work	95	9%
Other	177	17%
Total	1010	100%

Table 2: Distribution of the texts by topic.

3.2. Manual Annotation

The whole dataset has been annotated manually by four trained annotators[7], who have spent around three weeks training on the subject and discussing

[7]Following the recommendations of Basile et al. (2023), we provide more in-depth information about the annotators; we do this as disaggregated data to preserve their anonymity. All four are Italian native speakers, cisgender, and sensitive to gender-related issues due to either their studies, or their affinity to feminist movements and/or to the queer community, or their personal experience with gender-based discrimination. The annotation

Category	Example
ROLE	Cento uomini possono creare un accampamento, ma serve una donna per fare una casa. ENG: *A hundred men can build a camp, but it takes a woman to make a home.*
PERSONALITY	Sentivo qualcosa di speciale e sai, una donna non sbaglia mai le sensazioni. ENG: *I felt something special and you know, a woman never gets her feelings wrong.*
COMPETENCE	[Commento ad articolo con titolo "La pilota della British Airways ubriaca in volo: cacciata dall'aereo, aggredisce pure i poliziotti"] Come si possono affidare le sorti di un aereo ad una donna? ....scherzo, naturalmente... ENG: *[Comment on an article titled "British Airways pilot drunk on flight: kicked off plane, she even attacks police"] How can you trust a plane's fate to a woman? ....just kidding, of course...*
PHYSICAL	Oppure c'hanno le 5^, vanno in giro scollate come i manifesti messi d'inverno, e poi se rimani ""attirato"" dalle loro protuberanze ci rimangono male Povere cucciole. ENG: *Or they are a size D, they walk around with low-cut tops like winter posters, and then if you get ""attracted"" by their protuberances, they get upset. Poor little things.* [In the Italian text, there's a pun in the word *scollate*, which can mean both *wearing a low-cut top* and *coming unstuck because the glue has worn out*.]
SEXUAL	[Rivolto a una utente donna] fatevi voi una doccia e copritevi. Le donne vere si coprono. Gli animali vanno in giro nudi. ENG: *[Addressed to a female user] Take a shower and cover up. Real women cover up. Animals go around naked.*
RELATIONAL	[Commento a meme con testo "Aspettavo che mi mandassi tu un messaggio" e sotto l'immagine di un uomo vestito da principessa] Tipico post da zitella ENG: *[Comment on a meme with the text "I was waiting for you to text me" and underneath a picture of a man dressed as a princess] Typical spinster post*

Table 3: Examples of texts belonging to the six GS categories.

and defining the annotation guidelines.[8] Both the extensive preparation phase[9] and the involvement of multiple trained annotators ensured a shared understanding of the task and consistency in the application of the criteria, thereby contributing to the overall quality and reliability of the dataset.

Furthermore, as one of the key contributions of this work, we propose a new taxonomy for the semantic classification of gender stereotypes, with each category representing a different dimension of this phenomenon. This classification, which is outlined in the annotation guidelines mentioned above, was developed to capture the variety of ways in which stereotypes manifest in language and to support both linguistic analysis and automatic detection tasks.

For each text, the following information is provided:

- **GS value**: a number in the interval $[0 - 1]$ indicating the degree to which the text reflects or refers to a gender stereotype (where 1 is the maximum and 0 is the minimum GS degree);
- **GS category**: the category to which the gender stereotype (if present) belongs to.

GS Value Annotation. The overall annotation procedure consists of two steps:

1. Each annotator manually assigns, for each short text, a binary label *yes*/*no* indicating whether or not the text reflects or refers to a GS;
2. The GS value is computed by combining the four individual annotations.

Although the dataset was annotated by four trained annotators, the inherent subjectivity of the task inevitably introduced a certain level of disagreement. Following the perspectivist approach we opted for merging all annotations into a numerical GS value, rather than selecting a binary label obtained through annotation aggregation on the basis of majority voting [10]. This choice aligns with recent findings which indicate that leveraging disagreement is more convenient than reliably trying to eliminate it (Basile et al., 2023; Muscato et al., 2024).

As shown in Table 1, the underlying assumption is that full inter-annotator agreement (IAA) corresponds to the endpoints of the continuum: if all four annotators agree that there is no GS, the resulting GS value is 0; if all four annotators agree that there is a GS indeed, then the resulting GS value is 1. On the other hand, disagreement between the annotators in the selection of the binary label is supposed to indicate intermediate GS values, such as 0.25 (three *no* labels and one *yes* label), 0.5 (two *yes* labels and two *no* labels), and 0.75 (three *yes* labels and one *no* label).

GS Value annotation and the strictly dependent Gender Stereotype Detection task was the main focus of the above-mentioned GSI:detect shared task.

GS Category Annotation. If a text is marked by any annotator as containing or referring to a GS, it is also assigned to one of the six GS categories,

team consists of three women and one man: two are between 20 and 30 years old, one between 30 and 40, and one above 40. In terms of education, one holds a PhD in linguistics and three hold a master's degree in either linguistics, foreign language studies or computational linguistics.

[8]The annotation guidelines are available for download at this link.

[9]Which includes the study of existing literature on GSs (see 2.1), preliminary annotation of a subset of the GSI:detect dataset and the discussion of disagreement.

[10]The distributed dataset contains both the non-aggregated annotations by the four annotators, and the merged numerical GS value.

according to the classification described in the annotation guidelines and summarised as follows (examples are provided in Table. 3):

- ROLE: social and cultural expectations about what women and men should do and about how they should be;
- PERSONALITY: emotional and behavioural traits assigned to men and women based on their gender;
- COMPETENCE: generalized judgments of a person's abilities based on their gender;
- PHYSICAL: expectations about the physical appearance of men and (especially) women, and all aspects of personal care in general;
- SEXUAL: attitude and behaviour that men and women should have regarding sexuality;
- RELATIONAL: the way in which women and men should behave in interpersonal/sentimental relations.

To the best of our knowledge, this is the first attempt to provide a systematic taxonomy to classify gender stereotypes into semantic categories; it is not intended to be fully exhaustive, however, as it is derived from an abstraction over the direct observation of the examples contained in our dataset.

Given the more explorative nature of this level of annotation, its derived Gender Stereotype Classification task was presented at Evalita 2006 as a pilot subtask. Accordingly, we opted for a more traditional approach where a category is selected based on majority voting in case of disagreement between the annotators, resorting to a super-judge in case of ties (this however amounted to only 60 out of 1,010 entries, approximately 6%). For the same reason, we also adopted some simplifications in the annotation guidelines, such as to always select only one category.

Inter-Annotator Agreement (IAA) The total IAA between the four annotators on the choice of the *yes* or *no* label is 0.613 (Fleiss' k), which is a moderate agreement. Figure 1 provides an overview of the inter-annotator agreement among the four annotators, visualized through pairwise Cohen's k values for both the GS value and GS category annotations. The two heatmaps respectively illustrate the agreement patterns for the numerical and categorical scoring schemes.

As shown in Figure 1a, A2 and A3 have the highest agreement (0.679), A1 and A4 have the lowest agreement (0.486). A1 has the lowest average pairwise IAA (0.579), followed by A4 (0.583), A3 (0.631) and A2 (0.659). Regarding the GS category annotation (Figure 1b), the four experts scored another moderate agreement, with a IAA of 0.611 (Fleiss' k). Inspecting again the pairwise IAA values (Cohen's k), A2 and A3 have again the highest agreement (0.684), while A1 and A4 have the lowest agreement (0.509). In this case, however, A4 is the one with the lowest average pairwise IAA (0.573), followed by A1 (0.587), A3 (0.622) and A2 (0.657).

	Dev set	Test set	Total
WITH CONTEXT texts	82	323	405
NO CONTEXT texts	118	487	605
Total	200	810	1,010

Table 4: Dataset's size and split.

	Tokens	Items	Av. length
Texts only	33,673	1,010	33.3 tok.
Contexts only	18,445	405	45.5 tok.
Whole dataset	52,118	1,010	51.6 tok.

Table 5: Dataset's size in details.

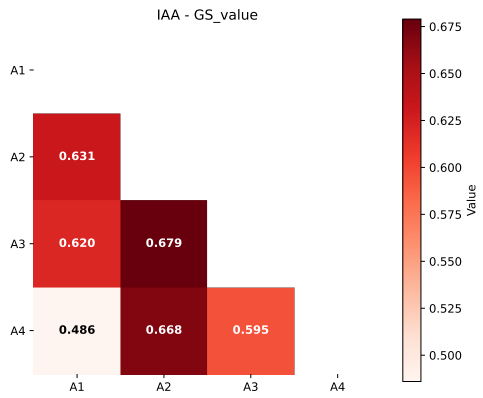
3.3. Statistics

As the dataset has been used in the context of a shared task, its 1,010 texts have been split as follows (Table 4): 20% of the dataset is used as development data (*dev set*), while the remaining 80% of the dataset is used for the official evaluation and ranking of participant systems (*test set*). The rationale behind this proportion is to balance the need for sufficient data for model tuning with the availability of a larger and more representative test set for evaluation purposes. Both the dev and the test set have a ratio of around 58% texts with context and 41% texts without context.

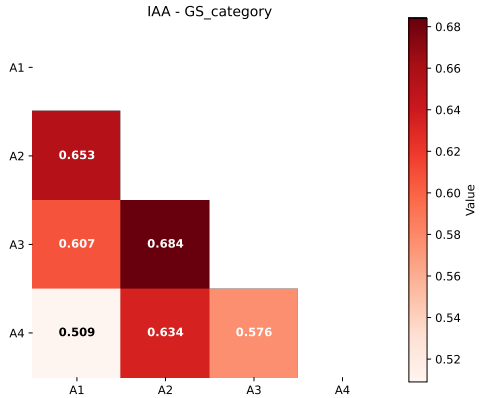
Table 5 reports detailed information about the size of the dataset in terms of tokens. The token count was computed using the Italian rule-based tokenizer included in the *spaCy* library[11] (version 3.8.7) as part of the *it_core_news_sm* linguistic model. The average length of the GSI:detect texts is 33.3 tokens if we consider pure original text and 51.6 in we include the added contextual information. The context metadata (added to 405 texts) consist on average of 45.5 tokens.

When defining the development and test splits, particular attention was paid to maintaining a balanced distribution of examples across both sets. As shown in Table 6, the overall proportion (i.e., the percentage) of items assigned to each GS value in the *Total %* column is approximately reflected in the relative composition of both the dev set and test set, when considering the ratio between the number of items per GS value and the total size of

[11] https://spacy.io



(a) IAA – GS_value



(b) IAA – GS_category

Figure 1: Comparison of IAA across the two annotation tasks, i.e. (a) GS value and (b) GS category annotation. Each heatmap visualizes pairwise agreement between annotators.

GS value	**Dev set**	**Test set**	**Total**	**Total%**
0	60	242	302	29.90%
0.25	25	84	109	10.79%
0.50	27	85	112	11.09%
0.75	25	105	130	12.87%
1	63	294	357	35.35%
	200	810	1,010	

Table 6: Dataset distribution by GS value.

Category	**Dev set**	**Test set**	**Total**	**Total%**
Role	30	107	137	13.56 %
Personality	29	108	137	13.56%
Competence	34	120	154	15.25%
Physical	20	90	110	10.89%
Sexual	14	72	86	8.52%
Relational	13	71	84	8.32%
GS value = 0	60	242	302	29.90%
	200	810	1,010	

Table 7: Dataset distribution by GS category.

the respective subset. This indicates that the split preserves the original distribution of GS values, ensuring a consistent representation of different degrees of stereotypicality across both subsets. A similar balance is maintained with the GS category (see Table 7), where the relative proportions of the six stereotype dimensions remain comparable between the dev and test sets.

This careful selection confirms that both subsets are representative of the overall dataset, providing a reliable basis for model tuning and evaluation, while avoiding unwanted biases in category distribution.

The current release of the dataset maintains the split between dev and test set. It also preserves the non-aggregated judgments of all annotators, thus allowing system to both learn from disagreement and test their predictions against the GS value emerging from all judgments.

4. Experiments

In order to provide reference baselines for future research, we conduct a series of experiments on the GSI:detect dataset. The aim of these experiments is not to achieve state-of-the-art performance, but rather to establish an initial benchmark showing how a predefined set of LLMs perform on the two tasks for which the dataset was designed: (i) the GS value detection and (ii) the GS category prediction.

Experimental setting. All models are evaluated in a zero-shot setting, without any fine-tuning, hyperparameter optimisation, or even task-specific adaptations. To ensure a fair and consistent evaluation, a minimal prompt engineering phase was carried out on the dev set: four researchers independently proposed one prompt each for both tasks (i.e., GS value detection and GS category prediction), and the best-performing prompt was selected based on preliminary experiments conducted with the *GPT5-nano model* (OpenAI, 2025). The final evaluation was then performed on the test set using the three selected models.

This experimental design allows us to provide transparent and reproducible baselines that can serve as a point of comparison for future studies exploring more advanced or fine-tuned approaches.

Tasks. We prompted the models instructing them to output a GS value between or equal to 0 and 1 (*GS value detection* task) and a GS category label for each instance in the dataset (*GS category prediction* task).

4.1. Models

To establish reference baselines on the GSI:detect dataset, we evaluate three LLMs differing in architecture, size, and accessibility, thereby including both closed-source models and open-source one.

GPT-5. (OpenAI, 2025) We use the *gpt-5-nano-2025-08-07* variant, a lightweight OpenAI model optimized for classification and reasoning tasks, employed in its text-only configuration.
GPT-4o. (OpenAI, 2024b,a) A proprietary multimodal model from the GPT-4 family, widely recognized for its strong (multimodal) reasoning and language understanding skills and used exclusively with its linguistic component.
Qwen3-14B. (Qwen-Team, 2025) An open-source transformer model available on the Hugging Face Hub. We employ the 14B-parameter variant.

Note that for all the models we tested both a *Split Prompt* configuration – where each task was addressed separately[12] – and a *Unified Prompt* configuration, where both tasks were handled within a single prompt.

4.2. Metrics

We adopted task-specific metrics to ensure a fair and accurate evaluation of model performance.

GS value Detection Task. The comparison of the models' performance on the GS value detection task is based on a normalized score derived from the Mean Squared Error (MSE), as reported in Table 8. MSE measures the average squared difference between the predictions and gold values, with larger errors more penalised, and is normalized by the variance of the data distribution to obtain NMSE. The final metric is defined as $\frac{1}{1+\text{NMSE}}$, bounded in $[0, 1]$, and is adopted for model comparison and ranking due to its improved interpretability. We compute also the Concordance Correlation Coefficient (CCC) score, that shows the agreement between predicted values and gold values, and how consistently the predictions align with gold values. CCC scores range from -1 (perfect disagreement) to $+1$ (perfect agreement), with higher CCC values indicating better model performance.
GS category Prediction Task. We assess the models' performance on the GS category prediction task using Macro F1 (not accounting for class imbalance) and Micro F1 scores (accounting for class imbalance), as seen in Table 9. A breakdown of models' per-category performance is provided in Figure 2, where higher F1 indicates better performance.

[12] i.e. with different prompts and different API calls.

For both tasks, models were also evaluated against several baselines.

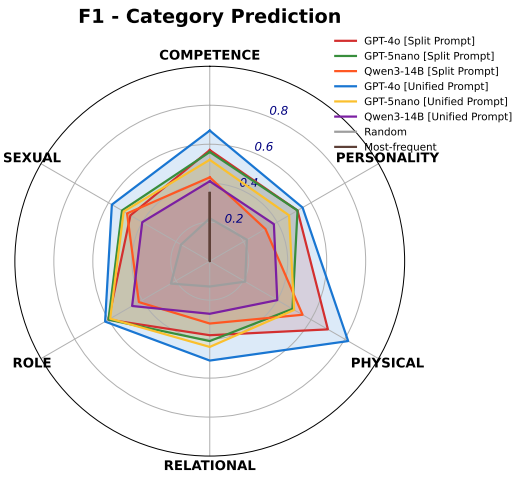


Figure 2: F1 score per Category

5. Results and Discussion

Table 8 presents results for the GS value detection task, clearly showing GPT-5-nano achieving the best overall performance in the split-prompt setting, with the highest normalized score (0.61) and CCC (0.60). GPT-4o exhibits a comparable behavior in the unified prompt setting with respect to the error-based metric, although with lower agreement. While the unified prompt setting does not bring substantial improvements for GPT-5-nano , it appears to benefit GPT-4o, exhibiting an increase of 0.10 points. Qwen3-14B does not achieve competitive performance overall. While in the unified prompt setting it reaches 0.54, slightly surpassing GPT-4o in the split configuration, it remains the weakest model in the split-prompt setting (0.49), where it achieves the lowest score among all evaluated models. Overall, all evaluated models consistently outperform both random and worst-case baseline. The same holds when considering the *0.5 baseline*[13], which assigns the midpoint value to all instances. However, Qwen3-14B in the split-prompt configuration does not surpass this baseline (0.49), making it the weakest model overall across both prompt configurations.

GS category prediction task results are described in Table 9. GPT-4o achieves the best Macro F1 score in the split-prompt configuration (0.54), and a clear improvement in Micro F1 (0.63) within the unified one. This indicates better handling of class imbalance. By contrast, Qwen3-14B consistently performs poorly also in this task,

[13] Results of this baseline perfectly correspond to 0.5 due to the balanced distribution of the dataset and the fact that GS values lie in the interval $[0, 1]$.

Setting	Model	1/(1+NMSE) ↑	CCC ↑
Baselines	Random	0.40	0.00
	Worst-case	0.18	-0.87
	0.5 value	0.50	0.00
Split Prompt	GPT-4o	0.51	0.43
	GPT-5nano	**0.61**	**0.60**
	Qwen3-14B	0.49	0.38
Unified Prompt	GPT-4o	0.61	0.55
	GPT-5nano	0.59	0.57
	Qwen3-14B	0.54	0.46

Table 8: GS value detection results. Best performance in bold.

Setting	Model	Macro F1 ↑	Micro F1 ↑
Baselines	Random	0.19	0.20
	Most-frequent	0.06	0.21
Split Prompt	GPT-4o	**0.54**	0.54
	GPT-5nano	0.52	0.53
	Qwen3-14B	0.38	0.40
Unified Prompt	GPT-4o	0.54	**0.63**
	GPT-5nano	0.50	0.52
	Qwen3-14B	0.39	0.40

Table 9: GS category prediction results for the evaluated models in a zero-shot setting. Best performance in bold.

achieving the lowest results among all evaluated models in both configurations (0.38 Macro F1 and 0.40 Micro F1 in the split setting), with only marginal differences between split and unified prompts (0.38 vs 0.39 Macro F1). Notably, all models remain substantially above both baselines, namely the random assignment and the most-frequent strategy (which always predicts the most common class, i.e., *competence*). Finally, Figure 2 (F1 scores per categories) shows that most models perform similarly across categories. However, the GPT-4o model not only outperforms the others, especially in the Unified Prompt setting, but shows a particular expertise in the PHYSICAL GS dimension.

Thus, overall, GPT-based models outperform the open model, with GPT-5-nano and GPT-4o respectively excelling in task 1 and 2. Such results highlight several interesting trends across the two tasks: GPT-5-nano is particularly aligned to humans capturing the gradient and continuous nature of stereotypicality in language, in contrast, GPT-4o performs better in the categorical classification of stereotypes. Moreover, the poor performance of Qwen3-14B across both tasks may be attributed to its smaller scale or to a training set of data that lacks the right amount of exposure to Italian data and gender-related social phenomena, suggesting that such performance may be also influenced by the ability of the model to capture language- and culture-specific patterns.

Finally, while almost all models outperform the random, worst-case and 0.5 baselines, the improvements over them remain relatively limited. In particular, when considering the 0.5 baseline, even the best-performing model (GPT-5-nano in the split setting) exceeds it by only 0.11 points, with GPT-4o showing a marginal gain of 0.01 points in the same configuration and a larger improvement (0.11) in the unified setting, and Qwen3-14B improving by only 0.04 points in its best configuration. Combined with the modest improvement over the random reference (~20 points for GPT-based models and ~10 for Qwen3-14B), this suggests that GS value prediction and, more generally, the assignment of reliable continuous scores remains a non-trivial task even for state-of-the-art systems.

6. Conclusions and Future Work

We presented GSI:detect, a new Italian resource for studying gender stereotypes in Italian short texts. The dataset introduces several innovations: (i) it includes cases of Gender Stereotype in naturally occurring contexts that goes beyond hate speech, (ii) it applies a perspectivist annotation that values disagreement, and (iii) it proposes for the first time a fine-grained taxonomy of gender stereotype categories. Experiments with LLMs show that the closed-source models considered in this study align more closely with human judgments, both in detecting the degree of stereotypicality and in categorical classification. However, given that only one relatively small open-source model (i.e, *Qwen3-14B*) was evaluated, and that it may have had more limited exposure to Italian data and its culturally grounded phenomena during training or instruction tuning, this result should be interpreted with caution and not generalized to all open-source systems, but rather viewed as an initial exploratory comparison in this direction. In addition, the relatively limited margin over the random baseline suggests that modeling the graded nature of stereotypicality remains challenging even for state-of-the-art (closed-source) systems.

Given the good response obtained by the participants in the shared task we organised based on GSI:detect, as future work we plan to consolidate our work by releasing a new version of our dataset applying the perspectivist approach to the classification of GSs. In addition to this, future work will focus on the sociolinguistic profiling of LLMs to better understand how their behaviour towards this topic aligns – or diverges – from human perspectives, with special attention to distinct demographic groups.

7. Acknowledgements

This paper is the result of a collaboration between all authors. Specifically, Davide Testa wrote Section 1, Section 5 and Section 6; Sofia Brenna wrote Section 4 together with Davide Testa; Manuela Speranza wrote Section 3 together with Gloria Comandini who also wrote Section 2.
This work has been carried out while Davide Testa was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Fondazione Bruno Kessler (FBK). Bernardo Magnini was supported by the PNRR MUR project PE0000013-FAIR (Spoke 2).

8. Ethical considerations

The GSI:detect dataset includes naturally occurring texts that may contain sexist or offensive content, collected solely for research purposes to study gender stereotypes. All data come from public sources and were anonymized to protect privacy. The views and opinions expressed in the dataset do not necessarily reflect those of the authors, and in some instances neither those of the informants, as the authors may have selected only an extract from the original texts.

9. Bibliographical References

L. Arcuri and M.R. Cadinu. 1998. *Gli Stereotipi. Dinamiche psicologiche e contesto delle relazioni sociali*. Il Mulino, Bologna.

V. Basile, F. Cabitza, and A. Campagner. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Toward a Perspectivist Turn in Ground Truthing for Predictive Computing*, pages 6860–6868, Washington DC. Association for the Advancement of Artificial Intelligence.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of preaggregation on the evaluation of highly subjective tasks. In *Proceedings of the AIxIA 2020 Discussion Papers Workshop*, pages 31–40. CEUR Workshop Proceedings.

Irene Biemmi. 2020. *Educazione sessista. Stereotipi di genere nei libri delle elementari*. Rosenberg & Sellier, Torino.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theorygrounded measurement of u.s. social stereotypes in english language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1276–1295, Seattle. Association for Computational Linguistics.

S. Cavagnoli and F. Dragotto. 2021. *Sessismo*. Mondadori, Milano.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective isms? on the danger of conflating hate and offence in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, page 275–282. Association for Computational Linguistics.

Gloria Comandini, Manuela Speranza, Sofia Brenna, Davide Testa, Stefania Cavagnoli, and Bernardo Magnini. 2026. Gsi:detect at evalita 2026: Overview of the task on detecting gender stereotypes in italian. In *Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026)*, Bari, Italy. CEUR.org.

E. Fersini, D. Nozza, and P. Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA Evaluation of NLP and Speech Tools for Italian. Proceedings of the Final Workshop*, pages 59–66, Torino. Accademia University Press.

C. Harrison and K. D. Tanner. 2018. Language matters: Considering microaggressions in science. *CBE - Life Sciences Education*, 17:1–8.

Lynne M. Jackson. 2011. The psychology of prejudice: From attitudes to social action. *American Psychological Association*.

H. Kirk, W. Yin, B. Vidgen, and P. Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, page 2193–2210, Stroudsbourg. Association for Computational Linguistics.

Manfred Klenner, Anne G¨ohring, and Michael Amsler. 2020. Harmonization sometimes harms. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. CEUR Workshop Proceedings.

Marco Madeddu, Simona Frenda, Mirko Lai, Viviana Patti, and Valerio Basile. 2023. Disaggreghate it corpus: A disaggregated italian dataset of hate speech. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 243–250. CEUR Workshop Proceedings.

M. Mitchell, G. Attanasio, I. Baldini, M. Clinciu, J. Clive, P. Delobelle, M. Dey, S. Hamilton, T. Dill, J. Doughman, R. Dutt, A. Ghosh, J. Zosa Forde, C. Holtermann, L.A. Kaffee, T. Laud, A. Lauscher, R.L. Lopez-Davila, M. Masoud, N. Nangia, A. Ovalle, G. Pistilli, D. Radev, B. Savoldi, V. Raheja, J. Qin, E. Ploeger, A. Subramonian, K. Dhole, K. Sun, A. Djanibekov, J. Mansurov, K. Yin, E. Villa Cueva, S. Mukherjee, J. Huang, X. Shen, J. Gala, H. Al-Ali, T. Djanibekov, N. Mukhituly, S. Nie, S. Sharma, K. Stanczak, E. Szczechla, T. Timponi Torrent, D. Tunuguntla, M. Viridiano, O. Van Der Wal, A. Yakefu, A. Névéol, M. Zhang, S. Zink, and Z. Talat. 2025. Shades: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11995–12041, Kerrville. Association for Computational Linguistics.

Benedetta Muscato, Chandana Sree Mala, Marta Marchiori Manerba, Gizem Gezici, and Fosca Giannotti. 2024. An overview of recent approaches to enable diversity in large language models through aligning with human perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, page 49–55. ELRA and ICCL.

OpenAI. 2024a. Gpt-4 technical report.

OpenAI. 2024b. Gpt-4o system card.

OpenAI. 2025. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf. Accessed: October 2025.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1246–1266, New York. Association for Computing Machinery.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

G. Peruzzi, V. Bernardini, and R. Lombardi. 2019a. Le questioni di genere dei giovani. un'indagine sulle percezioni e le esperienze di studenti e studentesse universitari. In G. Peruzzi, V. Bernardini, R. Lombardi, C. Rinaldi, M. Bacio, L. Bainotti, and G. Viggiani, editors, *Il bias del gender*, pages 13–50. Durango, Andria.

G. Peruzzi, V. Bernardini, R. Lombardi, C. Rinaldi, M. Bacio, L. Bainotti, and G. Viggiani. 2019b. *Il bias del gender*. Durango, Andria.

L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso. 2023. Overview of exist 2023: sexism identification in social networks. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, pages 593–599, Cham. Springer.

Qwen-Team. 2025. Qwen3 technical report.

G. Rizos and B.J. Schuller. 2020. Average jane, where art thou? – recent avenues in efficient machine learning under subjectivity uncertainty. In M.J. Lesot, S. Vieira, M.Z. Reformat, J.P. Carvalho, A. Wilbik, B. Bouchon-Meunier, and R.R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 42–55. Springer, Cham.

M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and Stranisci M. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2798–2805, Paris. European Language Resources Association (ELRA).

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

D.W. Sue. 2010. *Microaggressions in everyday life. Race, gender and sexual orientation*. John Wiley & Sons, Hoboken.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 769–779, Seattle. Association for Computational Linguistics.

M. Wojatzki, T. Horsmann, D. Gold, and T. Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgement. In *Proceedings of the 14th conference on Natural Language Processing (KONVENS 2018)*, pages 110–120.