# Subjectivity in Stereotypes Against Migrants in Italian: An Experimental Annotation Procedure

Soda Marem Lo[1,*,†], Marco A. Stranisci[1,2,*,†], Alessandra Teresa Cignarella[2,3], Simona Frenda[2,4], Valerio Basile[1], Cristina Bosco[1], Elisabetta Jezek[5] and Viviana Patti[1]

[1] *Università di Torino, Dipartimento di Informatica, Corso Svizzera 185 – 10149 Turin, Italy*
[2] *aequa-tech, Via Quarello 15/A – 10153 Turin, Italy*
[3] *Ghent University, Language and Translation Technology Team, Groot-Brittanniëlaan 45 – 9000 Ghent, Belgium*
[4] *Interaction Lab, Heriot-Watt University, EH14 4AS Edinburgh, Scotland*
[5] *Università di Pavia, Department of Humanities, Piazza del Lino 2 – 27100 Pavia, Italy*

**Abstract**

The presence of social stereotypes in NLP resources is an emerging topic that challenges traditionally used approaches for the creation of corpora and resources. An increasing number of scholars proposed strategies for considering annotators' subjectivity in order to reduce such bias both in computational resources and in NLP models. In this paper, we present Open-Stereotype, an annotated corpus of Italian tweets and news headlines regarding immigration in Italy developed through an experimental procedure for the annotation of stereotypes aimed to investigate their different interpretation. The annotation is the result of a six-step process, where annotators identify text-spans expressing stereotypes, generate rationales about these spans and group them in a more comprehensive set of labels. Results show that humans exhibit high subjectivity in conceptualizing this phenomenon, and that the prior knowledge of an Italian LLM leads to more consistent classifications of specific labels that do not depend on annotators' background.

**Keywords**

Subjectivity, Annotation, Italian, Stereotypes, Social Bias

## 1. Introduction

Developing fair Natural Language Processing (NLP) technologies for the detection of abusive language is still nowadays an open issue that gathers the attention of many scholars. The increasing awareness that corpora for hate speech detection exhibit significant biases, particularly favoring Western and white populations [1], has led scholars to foster explainability [2, 3] and cultural representativeness [4, 5] in the design of new resources. Furthermore, the growing number of perspectivist [6, 7] and multilingual [8] datasets contributes to a deeper and culturally aware understanding of abusive language, paving the way for the development of less biased technologies.

Recently, specific attention has been paid in particular to the presence of stereotypes in different contexts, such as political discourse [9], reactions to fake news [10], news comments [11], news and social media messages [12, 13] often through the development of taxonomies and annotated corpora. However, these advances do not encompass the diverse perceptions or interpretations of stereotypes in the text. For instance, despite some corpora for the detection of origins-related stereotypes have already been released [12, 14, 11, 15, 16], to the best of our knowledge, only one of them has been designed to take into account **subjectivity** [17] presenting the annotation of three different annotators. This limitation intersects with the scarcity of studies on bias and disagreement in the design of annotation schemes [18, 19, 20].

In this work we address this research gap by presenting the **Open Stereotype (O-Ster)[1] corpus**: a sub-portion of 1,022 texts of the HaSpeeDe corpus [12] (see details in Section 3) newly re-annotated through an experimental annotation procedure in which labels are not defined *a priori*, but they are rather defined throughout the annotation process highlighting annotator subjectivity about stereotypes (*a posteriori*). The resulting annotated corpus allowed us to reply to the following research questions:

• **(RQ 1). How do annotators recognize and conceptualize stereotypes?** We designed an annotation procedure that provides the identification of textual spans

*CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy.*

* Corresponding authors.
† These authors contributed equally.
sodamarem.lo@unito.it (S. M. Lo); marcoantonio.stranisci@unito.it (M. A. Stranisci); alessandrateresa.cignarella@ugent.be (A. T. Cignarella); s.frenda@hw.ac.uk (S. Frenda); valerio.basile@unito.it (V. Basile); cristina.bosco@unito.it (C. Bosco); elisabetta.jezek@unipv.it (E. Jezek); viviana.patti@unito.it (V. Patti)
0000-0002-5810-0093 (S. M. Lo); 0000-0001-9337-7250 (M. A. Stranisci); 0000-0002-4409-6679 (A. T. Cignarella); 0000-0002-6215-3374 (S. Frenda); 0000-0001-8110-6832 (V. Basile); 0000-0002-8857-4484 (C. Bosco); 0000-0003-2518-5200 (E. Jezek); 0000-0001-5991-370X (V. Patti)



[1] https://github.com/SodaMaremLo/Open-Stereotype-corpus.

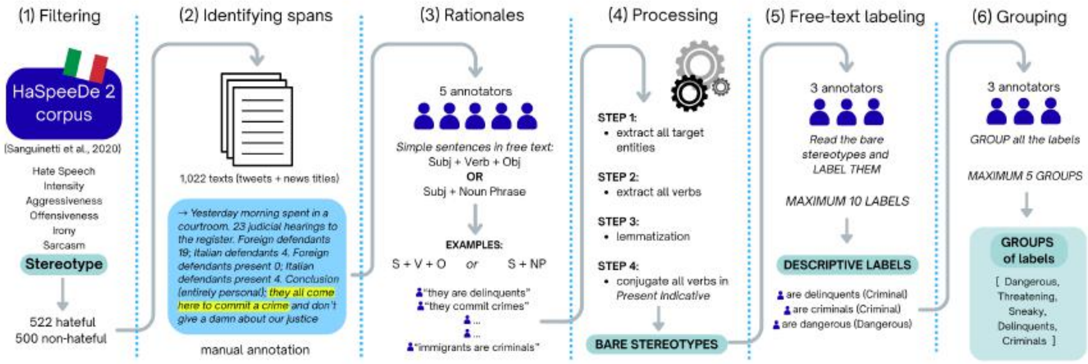


**Figure 1:** Visual representation of the full procedure employed for data filtering and annotation.

expressing stereotypes, the open-ended generation of rationales about their choice, and the categorization of rationales within a closed set of labels. The procedure showed how stereotypes in the same texts are differently perceived by humans, leading to the categorization of the same expressions in different and creative ways that might depend on the subjectivity of annotators.

• **(RQ 2). How do models conceptualize stereotypes?** In this first study, we prompted one specific Large Language Model (LLM), i.e., Minerva [21], to generate labels to categorize stereotypes. Observing which labels were created and with which annotator they agreed most of the time, we noticed that the LLM aligns more with the labels *Exploiters*, *Dangerous* and *Protected*, choosing them consistently throughout different classification runs.

## 2. Related Work

The detection and modeling of stereotypes in NLP has gained increasing attention in recent years, particularly as the field moves toward more socially responsible and inclusive language technologies. While early computational approaches primarily focused on gender bias and hate speech [22, 23], new work has begun to explore the broader phenomenon of stereotypes, including their implicit [24] and explicit manifestations across different social groups and languages [25, 26].

Most current work emphasizes the importance of distinguishing between stereotypes, prejudice, and discrimination, and highlights the advantages of a more interdisciplinary approach between computational linguistics and social psychology [27]. The Stereotype Content Model (SCM) [28] and its extension, the ABC model [29], have been quite often adopted by NLP scholars to conceptualize stereotypes along dimensions such as warmth, competence, and belief alignment. These frameworks have informed both annotation schemes and computational models, enabling more structured analyses of stereotype content. Examples of their application are the work of Bosco et al. [25] and Schmeisser-Nieto et al. [14] in which the authors apply an SCM-based scheme for describing stereotypes towards migrants to a trilingual corpus of tweets.

Concerning Italian, the *HaSpeeDe2* shared task [12] was one of the first to explicitly address stereotype detection by means of a dedicated subtask. Results pioneered the way for research into stereotype detection in Italian social media, investigating the connection between hate speech and stereotypical content in models. Furthermore, the results of the shared task suggest the need to approach stereotype detection as a subtle and independent phenomenon from hate speech. Schmeisser-Nieto et al. [30] comparing the human annotation and model predictions on stereotype detection noted that models tend to show low confidence when annotators have more disagreement with each other, highlighting the importance of encoding plural interpretations in resources and models. In such context, Cignarella et al. [31] developed the QUEEREOTYPES corpus, in which annotator perspectives are encoded in labeling stereotypes towards LGBTQIA+ people.

Perspectives of annotators matter and studies such as those of Sap et al. [5] and Xia et al. [32], for instance, have shown that demographic factors such as ethnicity or personal and/or linguistic background, can significantly influence the perception of hate speech and stereotypes.

The present work builds on the key concepts outlined in this section, by **proposing an experimental annotation procedure** that (i) elevates annotator subjectivity and (ii) builds on narrative patterns in free-text descriptions of stereotypes against migrants. Rather than enforcing a harmonized gold standard, we create and release non-harmonized annotations to preserve the diversity of annotator perspectives.[2] This approach aligns with emerging best practices in participatory NLP, and contributes to the growing body of resources for stereotype detection—particularly in languages other than English.

## 3. Annotation Procedure

For the creation of the O-Ster corpus, we adopted a descriptive annotation scheme as previously done by Röttger et al. [19], with the overarching goal of emphasizing the **subjectivity of annotators** in recognizing and describing the presence of stereotypes in texts. The annotation procedure is composed of several steps as shown in Figure 1. In this section, we describe all the steps in detail.

**(1) Filtering the HaSpeeDe2 corpus.**
The annotation process began with the extraction of a specific subset from the HaSpeeDe2 dataset [12]. This dataset, originally annotated with the presence/absence of hate speech and stereotypes, has been extended also in other works with the annotation of various dimensions of harmful language, including Intensity, Aggressiveness, Offensiveness, Irony, and Sarcasm [33].

For our purposes, we focused on the subset of texts annotated with a stereotype value of 1. This filtered corpus consists of 1,022 tweets and news headlines, each explicitly marked as containing stereotypical content (of these, 522 texts are hateful and 500 are non-hateful).

**(2) Identification of textual spans.**
Five different annotators (all researchers in NLP) were instructed to identify one or more spans of text that explicitly conveyed stereotypical content. The annotation task was carried out using a simple spreadsheet, where annotators copied and pasted the identified spans into a designated column corresponding to each text entry and partially relied on the Label Studio[3] platform.

**(3) Writing of rationales.**
For each identified textual span, annotators were asked to provide a corresponding rationale that explicitly expresses the sense behind the stereotype and the targeted group. They should be provided in the form of a simple sentence, typically following either a Subject-Verb-Object (S-V-O) or Subject-Noun Phrase (S-NP) structure. Examples include: *"i rom hanno invaso l'Italia"*[4] (S-V-O) and *"gli immigrati sono privilegiati"*[5] (S-NP). This step resulted in a total of 3,578 *span–rationale pairs*.

**(4) Text processing.**
We processed the rationales to ensure consistency and facilitate further linguistic analysis. In particular, 1) we extracted all the target entities mentioned in the sentences; 2) we identified the verbs associated with the targets; 3) we applied lemmatization to reduce verbs to their base forms; and 4) conjugated them in the *Present Indicative* tense. This normalization step allowed us to reduce the rationales to a set of 576 *distinct bare stereotypes*. Finally, all rationales that appeared only once in the corpus were removed to ensure focus on recurring patterns, resulting in a total of 248 *frequently occurring bare stereotypes*.

**(5) Free-text labeling.**
To further consolidate the subset of *bare stereotypes* resulting from the previous step of the procedure into a manageable and interpretable taxonomy, three annotators were independently tasked with grouping them by generating 10 descriptive labels. Each label was designed to capture the underlying theme or semantic core shared by multiple rationales. For example, the statements "(they) are delinquents" and "(they) are criminals" might have been grouped under the descriptive label CRIMINAL, while "they are dangerous" might have been categorized under the descriptive label DANGEROUS. This process allowed the transformation of free-text rationales into a structured set of stereotype categories suitable for classification tasks.

**(6) Grouping.**
To reach a narrower level of the taxonomy, we asked the 3

| |
|---|
| **1. Text**: "Mattinata di ieri passata in un'aula di tribunale. 23 udienze al ruolo. Imputati stranieri 19; imputati italiani 4. Imputati stranieri presenti 0; imputati italiani presenti 4. Conclusione (del tutto personale); vengono tutti a delinquere qua e se ne fregano della nostra giustizia" TRANSLATION → *Yesterday morning spent in a courtroom. 23 judicial hearings to the register. Foreign defendants 19; Italian defendants 4. Foreign defendants present 0; Italian defendants present 4. Conclusion (entirely personal); they all come here to commit a crime and don't give a damn about our justice* |
| **2. Textual span**: *they all come here to commit a crime* |
| **3. Rationale (S-V-O)**: *[foreigners are delinquents, foreigners commit crimes, ..., immigrants are sneaky, ..., immigrants are criminals]* |
| **4a. Targeted entity**: foreign defendants = foreigners |
| **4b. Bare stereotype**: *[are dangerous, are threatening, are delinquents, are criminals]* |
| **5. Descriptive label**: *are delinquents* → Criminal |
| **6. Group**: *THREAT* |

**Table 1**
Example of an annotated text from the O-Ster corpus.

[2] We also include a positionality statement in Appendix A.1.
[3] https://labelstud.io/.
[4] Roma people invaded Italy.
[5] Migrants are privileged.

annotators to reduce the initial set of 10 descriptive labels to 5 broader groups. This second round of refinement involved merging semantically related labels to enhance clarity and usability. For example, the rationales "(they) are delinquents" and "(they) are criminals", previously grouped under the descriptive label CRIMINAL, and "they are dangerous", categorized under DANGEROUS, could all be further consolidated under the broader group THREAT. It is important to emphasize that, throughout the entire annotation process, annotators were given minimal (if any) prescriptive instructions. They received very limited annotation guidelines, which allowed for a more open-ended and subjective interpretation of stereotype groupings. This deliberate lack of constraints is a central feature of our experimental design, aimed at capturing the annotators' intuitive understanding (and subjectivity) of stereotypical content in Italian texts.

An example of a fully annotated text, including its associated stereotype and final label, is presented in Table 1 to complement the information of the workflow of the annotation procedure already outlined in Figure 1.

## 4. Corpus Analysis

O-Ster consists of $1,022$ texts annotated by 5 people in different proportions (Table 2). Almost all posts were annotated by two people, except for 27 by just one person. For each text, the annotator could assign multiple rationales, reaching an average of $1.77$ per post, and a total of $3,578$ annotations.

| Annotator | Nickname | #Texts | #Annotations |
|---|---|---|---|
| _01 | Duck | 747 | 1,367 |
| _02 | Bear | 75 | 112 |
| _03 | Lion | 100 | 129 |
| _04 | Panda | 94 | 178 |
| _05 | Rhino | 1,001 | 1,792 |

**Table 2**
Number of texts and annotations across each annotator. To anonymize and simplify references to annotators throughout this work, we chose arbitrary animal-themed nicknames to be used instead of numerical identifiers. These names are chosen solely for ease of reading and do not imply any characteristics of the annotators.

**Identifying 'agents' and 'patients' in the rationales.** From the third step described in Section 3, a total of $1,547$ rationales was reached. To better understand their construction, we looked into the role of the subject in terms of agents and patients. Specifically, we syntactically parsed each rationale and assigned the role of 'agent' to all the targets that are the subject of active verbs (*Migrants are criminals*), and 'patient' when they are the object of the sentence or the subject of a passive verb (*Migrants must be kicked out*). Finally, we performed a manual aggregation of Roma and Sinti in a unique category, as well as politicians including specific people and parties, and ethnic minorities named by referring to their origin, or with generic terms such as "foreigners".

Considering the unbalanced number and type of annotations across annotators, we computed the proportion of times each target was annotated as an agent (or patient) by each annotator. This was done by dividing the frequency of each target (as agent or patient) by the total number of agent or patient annotations made by that annotator. We then calculated per-annotator averages of these proportions to establish individual thresholds, used to highlight the most frequently annotated targets. Results are presented in Table 3.

Results show that for all annotators when targets are presented as *immigrant*, they tend to be framed as both agents and patients in high percentages. However, Bear and Rhino often give agency to specific ethnic minorities. When Italians are targets, they only play the role of agents, especially presenting rationales linked to financial supports, such as *Italians pay for immigrants*. Interestingly, Roma and Sinti are framed as patients by Duck, especially using the rationale *Roma are treated better than Italians*, and in a low percentage by Rhino (3.2%). Other annotators' rationales present them only as agents, more often as criminals.

| Target | Agency | Annotator | Frequency |
|---|---|---|---|
| Immigrants | Agent | Duck | 41.43% |
| Immigrants | Agent | Bear | 62.22% |
| Immigrants | Agent | Lion | 41.82% |
| Immigrants | Agent | Panda | 54.78% |
| Immigrants | Agent | Rhino | 40.6% |
| Italians | Agent | Bear | 7.78% |
| Italians | Agent | Panda | 12.74% |
| Ethnic minority | Agent | Bear | 15.56% |
| Ethnic minority | Agent | Rhino | 10.97% |
| Islamic | Agent | Duck | 13.14% |
| Islamic | Agent | Lion | 48.18% |
| Islamic | Agent | Panda | 12.1% |
| Islamic | Agent | Rhino | 9.9% |
| Roma and Sinti | Agent | Duck | 32.61% |
| Roma and Sinti | Agent | Panda | 10.19% |
| Roma and Sinti | Agent | Rhino | 31.41% |
| Immigrants | Patient | Duck | 61.38% |
| Immigrants | Patient | Bear | 57.14% |
| Immigrants | Patient | Lion | 91.67% |
| Immigrants | Patient | Panda | 50.0% |
| Immigrants | Patient | Rhino | 86.4% |
| Roma and Sinti | Patient | Duck | 19.31% |

**Table 3**
For each annotator, the table shows targets annotated as agents or patients whose frequency exceeds the annotator-specific threshold. Frequencies are reported as percentages, normalized within each annotator.

| Duck 10 C | Duck 5 C | Bear 10 C | Bear 5 C | Rhino 10 C | Rhino 5 C |
|---|---|---|---|---|---|
| Criminal | Subtle | Burden | Worsen our lives | Dangerous | Threat |
| Deceivers | Subtle | Invaders | Worsen our lives | Bullies | Threat |
| Burden | Parasites | Selfish | Do not contribute | Parasites | Exploiters |
| Privileged | Parasites | Loafers | Do not contribute | Invader | Exploiters |
| Dangerous | Incompatible | Dangerous | Dangerous | Lazy | Exploiters |
| Radicalized | Incompatible | Criminal | Dangerous | Radicalized | Radicalized |
| Problem | Problem | Degraded | Degraded | Worse than us | Ruin of Italy |
| Degraded | Immoral | Dirty | Degraded | Savage | Ruin of Italy |
| Bullies | Immoral | Different culture | Different culture | Degraded | Ruin of Italy |
| Uncivilized | Immoral | Different from us | Different culture | Protected | Protected |

**Table 4**
In grey the 10 descriptive labels, and in white the 5 grouped labels for each annotator. Duck corresponds to annotator_01, Bear to annotator_02, and Rhino to annotator_05.

**Label analysis.**

As described in Section 3, annotators were asked to group the bare stereotypes into 10 descriptive labels, and then categorize them in 5 broader groups. Results of these steps are presented in Table 4. Focusing on the ten descriptive labels (grey columns), it is possible to notice similarities across annotators. They all individuated the idea of dangerousness (*Dangerous*), referring to stereotypes connected to being violent. However, analysing the dataset, Duck characterises this description with the idea of invasion, Bear includes non-violent forms of dangers such as bringing diseases, while Rhino involves those aspects that the other two separated in the *Criminal* label, such as stealing and cheating.

Other similarities are in the idea of being degraded (*Degraded* by all annotators), lazy (*Loafers* by Bear, and *Lazy* by Rhino), and a burden (*Burden* by Duck and Bear, and *Parasites* by Rhino). The use of different words for similar concepts, already suggests the different focus adopted by each annotator. For example, *Loafers* was connected to being useless, more than simply acting as lazy.

Another interesting commonality is the idea of being backward people and also this concept is expressed through different labels across annotators. Duck used *Uncivilized*, Bear *Different culture*, while Rhino separated the concept into two descriptive labels: *Savage* and *Worst than us*.

Finally, some stereotypes have been labeled in significantly different ways. An example is *they are nomads*, assigned to *Privileged* by Duck, *Different from us* by Bear, and *Invader* by Rhino, highlighting people's fear of being conquered or having their territories squatted.

The way an annotator looks at a phenomenon and its categorization becomes even more evident when analyzing the last step: grouping the descriptive labels in 5 categories (white columns in Table 4). In fact, they are required to choose which concepts they believe to be priorities and capable of encompassing multiple stereotypes.

Duck does not connect the aspect of crime with the idea of danger, as might have been expected from looking at the choices of the other annotators (*Degraded* by Bear and *Threat* by Rhino). In contrast, *Criminal* was merged with *Deceivers*, **combining the dimension of crime with cheating**, and tagging the group as *Subtle*. On the other hand, *Dangerous* has been included with *Radicalized* in **the broader imagery of incompatibility**, implicitly defining what "we" is not. Bear's groups better encapsulate a contrast us *vs.* them, specifically with the labels *Worsen our lives* and *Different culture*, which concentrate in a single label **the aspects of diversity**, primarily religious and cultural. It is noteworthy how the annotators' positionality (Appendix A.1), in this case, is most evident through their clear-cut distinction between us and them—a trait that is often absent in Rhino's labels.

Both Duck and Rhino group the idea of being respectively uncivilized and savage with being degraded, the former using the expression *Immoral*, thus framing the three descriptive labels into a **moral stand**; the latter choosing *Ruin of Italy*, referring to the **effect of those acts**. Finally, *Exploiters* unifies the dimension of being parasites and lazy, with that of invasion, in a very broad group that defines exploitation from an economic and territorial point of view. Overall, there is a general focus on the exploitation of the country and of the caused sense of danger (respectively *Parasites* and *Subtle* by Duck, *Do not contribute* and *Dangerous* by Bear, and *Exploiters* and *Threat* by Rhino).

Each annotator, however, has elements of uniqueness. For Duck, this is reflected in the creation of a single group of stereotypes that define **aspects of the target groups' identities** perceived as problematic (*Problem*). Bear, on the other hand, is the only one to foreground the **idea of a worsening of Italians' lives**, defined in relation to the risk of invasion and economic exploitation.

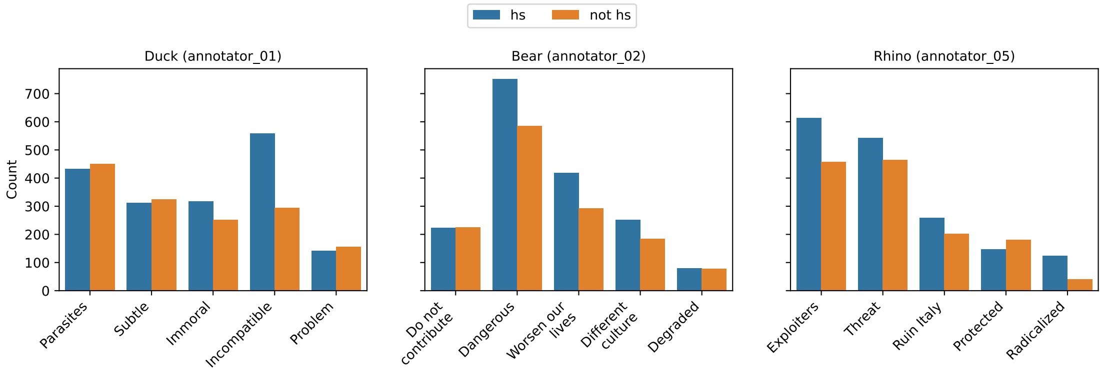


**Figure 2:** Count of each label occurrence for both hateful (blue) and not hateful (orange) texts, broken down by annotator. The labels are derived from the processes described in Section 3 and Table 4.

Lastly, Rhino is the only one to maintain a single label for the religious dimension and the perspective of protection, concerning the **perception of a privileged position** of the target group to Italians.

**Hateful comments.**

Focusing on the last phase of the pipeline, Figure 2 shows how the occurrence of the groups of labels changes based on the presence of hate speech. Labels such as *Incompatible*, *Dangerous*, and both *Exploiters* and *Radicalized* respectively for Duck, Bear and Rhino, tend to be more frequent when the message was annotated as hateful. These results highlight how a stereotypical representation of the stranger as an invader, religious extremist, or more generally a threatening individual, is linked to hate speech. It is worth noticing that the blue bars tend to be higher in most cases, although the texts are almost perfectly split across hateful and not-hateful (respectively 522 and 500). This indicates that the presence of hate speech also leads to the presence of multiple stereotypes in the same text.

## 5. Experiment

In this section, we present an experiment aimed at observing the behavior of an Italian LLM in the classification of stereotypes according to the labels derived from our annotation process (Section 3). The experimental setup was a zero-shot text generation task. We fed the LLM with a message and a list of the three labels defined by annotators and asked the model to generate as output one of the three labels.[6]

We repeated the experiment three times with three different randomizations of the order of the labels in the prompt, and used Minerva-7B-instruct-v1.0 to solve the task. On average, the model generated a bad output $7.32\%$ of the time. Messages that obtained a classification throughout all three runs are $1,922$: $85.61\%$ of the total. The analysis of results presented in this section considers only texts that obtained a classification in each run.

**Label distribution across runs.**

Given the high number of cases in which the LLM provides at least two different labels for the same text across the classification runs ($68.6\%$ of the time), we provided an analysis of group labels in runs when the LLM always produces the same output and when it always produces a different output. We considered two types of distributions: **Consistent** are the labels that are always predicted across the runs; **Inconsistent** are the labels produced in runs with at least one different prediction. In Table 5 the top-5 Consistent labels and the top-5 Inconsistent labels are reported. As can be observed, there are some labels that are more likely to be consistently predicted by the LLM across runs. It is the case of *Exploiters*, *Dangerous*, and *Protected*

| Stereotype Label | Annotation | Consistent | Inconsistent |
|---|---|---|---|
| Exploiters | 701 | 219 | 111 |
| Dangerous | 828 | 142 | 141 |
| Protected | 260 | 130 | - |
| Threat | 691 | 55 | 115 |
| Subtle | 387 | 19 | 90 |
| Incompatible | 552 | - | 93 |
| **Total** | **5,766** | **603** | **744** |

**Table 5**
The distribution of group labels in LLM's predictions that do not vary across runs (column 'Consistent') and predictions that are different in each run (column 'Inconsistent'). For each column, the absolute distribution of the top-5 labels is reported. In Column 'Annotation', the absolute distribution of group labels in the corpus is reported. The last row reports the total number of labels in each distribution.

[6] see Appendix A.2 for details about the prompt.

| Annotator | Group of label | Label distribution | run_1 | run_2 | run_3 |
|---|---|---|---|---|---|
| Duck | Parasites | 0.193 | 0.196 | 0.143 | 0.000 |
| | Immoral | 0.040 | 0.042 | 0.000 | 0.000 |
| | Incompatible | 0.375 | 0.379 | 0.143 | 1.0 |
| | Subtle | 0.363 | 0.371 | 0.143 | 0.000 |
| | Problem | 0.028 | 0.012 | 0.571 | 0.000 |
| Bear | Do not contribute | 0.056 | - | 0.071 | 0.055 |
| | Different culture | 0.064 | - | 0.286 | 0.051 |
| | Worsen our lives | 0.298 | - | 0.286 | 0.299 |
| | Degraded | 0.012 | - | 0.000 | 0.013 |
| | Dangerous | 0.568 | - | 0.357 | 0.581 |
| Rhino | Radicalized | 0.032 | 0.125 | 0.031 | 0.000 |
| | Exploiters | 0.448 | 0.500 | 0.432 | 0.692 |
| | Protected | 0.056 | 0.000 | 0.062 | 0.000 |
| | Threat | 0.464 | 0.375 | 0.476 | 0.308 |

**Table 6**
Distribution of labels assigned by the annotators and the model across the 248 comments where, in each run, the model selected a different annotator's label. Duck corresponds to annotator_01, Bear to annotator_02, and Rhino to annotator_05.

that combined represent 81.8% of the distribution. If the first two are the most occurring group labels defined by annotators (Section 3), *Protected* is not a common label, since it appears only 260 times in the corpus. This suggests that there is 50% chance that the LLM consistently predict the label *Protected* when encountering it, while the chance of having a consistent prediction of *Dangerous* is 17% and 31.2% for *Exploiters.* On the opposite side of this spectrum, there is *Threat*, which appears 691 times in the corpus but is consistently predicted only 55 times (7.9%). The distribution of labels predicted inconsistently by the LLM shows interesting results as well. There is a lower gap between most and less occurring labels among the top-5 (141 *versus* 93), suggesting that the model tends to spread inconsistent predictions among a more homogeneous pool of labels. *Dangerous* is the label that appears the most in LLM's inconsistent predictions, coherently with its distribution among the group labels in the corpus. *Threat* is the second-most occurring one, appearing in inconsistent predictions twice than consistent ones (115). This confirms the low ability of LLM to conceptualize this specific label. *Protected*, which is strongly present in consistent prediction, is not among the top-5 labels in inconsistent predictions, appearing only 14 times. Finally, it is worth mentioning that *Incompatible* appears 93 times in inconsistent predictions (third-most occurring) but only 3 times in consistent ones, suggesting that the LM struggles in the conceptualization of this group label as well.

**Consistent labels.**
As regards the Consistent labels, the model agreed across all the runs for a total of 603 annotations, selecting Rhino's labels 68.99% of the time, Bear's 26.37%, and Duck's 4.64%.

Considering the strong reliance on Rhino, we looked, in particular, at the labels generated by the model in this specific subset. Results show that it tends to prefer *Exploiters* and *Protected* over other annotators' labels, selecting both way more frequently than Rhino. Coherently with the previous analysis, *Exploiters* has a distribution of 0.365 by the human annotator *vs.* 0.526 by the model, while *Protected* respectively of 0.135 *vs.* 0.312. This shows that the reliance on Rhino should not be explained in terms of alignment to annotators' conceptualization of the stereotypes, but rather as a preference of the model towards this conceptualization. In fact, the other labels chosen by the same annotator rarely appear in this subset, with *Ruin of Italy* being totally missing.

**Inconsistent labels.**
Among the Inconsistent labels, we focused on cases where all runs disagree, resulting in 248 comments where the model chose a different annotator's label for each run. Table 6 presents humans' and models' label distribution on this specific subset. Results show that the model leans toward one annotator at a time, respectively Duck, Rhino and Bear for the first, second and third run. To further investigate this pattern, we checked whether the order of the variable, randomized for each run, had an influence on this result. We examined how often the selected label appeared in first position, and found that the annotator's label each run agrees with is almost always ranked first: specifically, 240, 227, and 234 times out of 248 for each of the three runs respectively. This highlights that when the model is less confident presents strong inconsistencies among the runs, and we infer it relies on the instruction example "Return as output (Output) a single option in the form of a Python list (e.g., ['Option 1'])"(Appendix A.2). These results necessitate a further analysis

of how LLMs manage challenging texts to annotate and low-confidence scenarios, which we plan to do in the future.

## 6. Conclusion and Future Work

In this paper, we presented O-Ster, a new corpus of Italian stereotypes annotated through an experimental framework. The corpus includes $1,022$ texts annotated at the span level. Each span has been complemented by a rationale expressing the individuated stereotype, and rationales served as a basis for the annotators to create labels associated with each text. This bottom-up process of label generation enabled observing how annotators with different backgrounds, and an LLM conceptualize the phenomenon. Results show a high subjectivity in the conceptualization of stereotypes by humans and the alignment of the LLM with certain specific labels in a zero-shot setting.

Future work will focus on expanding the corpus, in order to better understand how subjectivity affects this phenomenon and to what extent the annotation procedure may be generalizable and transferable to other languages and tasks of abusive language detection.

## Acknowledgments

The work of A. T. Cignarella is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions, Grant Agreement No. 101146287.

## References

[1] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1668–1678.

[2] T. Declerck, J. P. McCrae, M. Hartung, J. Gracia, C. Chiarcos, E. Montiel-Ponsoda, P. Cimiano, A. Revenko, R. Sauri, D. Lee, et al., Recent developments for the linguistic linked open data infrastructure, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 5660–5667.

[3] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, Semeval-2021 task 5: Toxic spans detection, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), 2021, pp. 59–69.

[4] S. H. Muhammad, I. Abdulmumin, A. A. Ayele, D. I. Adelani, I. S. Ahmad, S. M. Aliyu, N. O. Onyango, L. D. Wanzare, S. Rutunda, L. J. Aliyu, et al., Afrihate: A multilingual collection of hate speech and abusive language datasets for african languages, arXiv preprint arXiv:2501.08284 (2025).

[5] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, arXiv preprint arXiv:1911.03891 (2019).

[6] P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. Von Vacano, C. Kennedy, The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022, 2022, pp. 83–94.

[7] N. Lee, C. Jung, J. Myung, J. Jin, J. Camacho-Collados, J. Kim, A. Oh, Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 4205–4224.

[8] A. A. Monnar, J. Perez, B. Poblete, M. Saldaña, V. Proust, Resources for multilingual hate speech detection, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), 2022, pp. 122–130.

[9] J. J. Sánchez-Junquera, B. Chulvi, P. Rosso, S. P. Ponzetto, How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/8/3610. doi:10.3390/app11083610.

[10] E. Chierchiello, T. Bourgeade, G. Ricci, C. Bosco, F. D'Errico, Studying reactions to stereotypes in teenagers: an annotated Italian dataset, in: R. Kumar, A. K. Ojha, S. Malmasi, B. R. Chakravarthi, B. Lahiri, S. Singh, S. Ratan (Eds.), Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024, ELRA and ICCL, Torino, Italia, 2024, pp. 115–125. URL: https://aclanthology.org/2024.trac-1.13/.

[11] A. Ariza-Casabona, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, P. Rosso, Overview of detests at iberlef 2022: Detection and classification of racial stereotypes in spanish, Procesamiento del lenguaje natural 69 (2022) 217–228.

[12] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, C. Tommaso, V. Patti, R. Irene, HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task, in: EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and

Speech Tools for Italian, CEUR, 2020, pp. 1–9.
[13] P. Chiril, F. Benamara, V. Moriceau, "Be nice to your wife! The restaurants are closed": Can Gender Stereotype Detection Improve Sexism Classification?, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2833–2844.
[14] W. S. Schmeisser-Nieto, A. T. Cignarella, T. Bourgeade, S. Frenda, A. Ariza-Casabona, M. Laurent, P. G. Cicirelli, A. Marra, G. Corbelli, F. Benamara, et al., Stereohoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes, Language Resources and Evaluation (2024) 1–39.
[15] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5356–5371. URL: https://aclanthology.org/2021.acl-long.416/. doi:10.18653/v1/2021.acl-long.416.
[16] Z. Wu, S. Bulathwela, M. Pérez-Ortiz, A. S. Koshiyama, Stereotype detection in llms: A multiclass, explainable, and benchmark-driven approach, 2024. URL: https://api.semanticscholar.org/CorpusID:268856718.
[17] W. S. Schmeisser-Nieto, P. Pastells, S. Frenda, A. Ariza-Casabona, M. Farrús, P. Rosso, M. Taulé, Overview of detests-dis at iberlef 2024: Detection and classification of racial stereotypes in spanish-learning with disagreement, Procesamiento del Lenguaje Natural 73 (2024) 323–333.
[18] D. Hovy, S. Prabhumoye, Five sources of bias in natural language processing, Language and linguistics compass 15 (2021) e12432.
[19] P. Röttger, B. Vidgen, D. Hovy, J. Pierrehumbert, Two contrasting data annotation paradigms for subjective nlp tasks, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 175–190.
[20] A. Hautli-Janisz, E. Schad, C. Reed, Disagreement space in argument analysis, in: G. Abercrombie, V. Basile, S. Tonelli, V. Rieser, A. Uma (Eds.), Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 1–9. URL: https://aclanthology.org/2022.nlperspectives-1.1/.
[21] SapienzaNLP, sapienzanlp/minerva-7b-instruct-v1.0, https://huggingface.co/sapienzanlp/Minerva-7B-instruct-v1.0, 2024. Accessed in June 2025.
[22] K. Stanczak, I. Augenstein, A Survey on Gender Bias in Natural Language Processing, 2021. URL: http://arxiv.org/abs/2112.14168. doi:10.48550/arXiv.2112.14168, arXiv:2112.14168 [cs].
[23] P. Fortuna, S. Nunes, A Survey on Automatic Detection of Hate Speech in Text, ACM Computing Surveys 51 (2019) 1–30. URL: https://dl.acm.org/doi/10.1145/3232676. doi:10.1145/3232676.
[24] W. Schmeisser-Nieto, M. Nofre, M. Taulé, Criteria for the annotation of implicit stereotypes, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 753–762. URL: https://aclanthology.org/2022.lrec-1.80/.
[25] C. Bosco, V. Patti, S. Frenda, A. T. Cignarella, M. Paciello, F. D'Errico, Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP, Information Processing & Management 60 (2023) 103118. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306457322002199. doi:10.1016/j.ipm.2022.103118.
[26] T. Bourgeade, A. T. Cignarella, S. Frenda, M. Laurent, W. Schmeisser-Nieto, F. Benamara, C. Bosco, V. Moriceau, V. Patti, M. Taulé, A multilingual dataset of racial stereotypes in social media conversational threads, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 686–696. URL: https://aclanthology.org/2023.findings-eacl.51/. doi:10.18653/v1/2023.findings-eacl.51.
[27] A. T. Cignarella, A. Giachanou, E. Lefever, Stereotype Detection in Natural Language Processing, 2025. URL: https://arxiv.org/abs/2505.17642. arXiv:2505.17642.
[28] S. T. Fiske, A. J. C. Cuddy, P. Glick, J. Xu, A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition, Journal of Personality and Social Psychology (2002) 878–902.
[29] A. Koch, R. Imhoff, R. Dotsch, C. Unkelbach, H. Alves, The abc of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion., Journal of personality and social psychology 110 (2016) 675–709. URL: https://api.semanticscholar.org/CorpusID:6287638.
[30] W. S. Schmeisser-Nieto, P. Pastells, S. Frenda,

M. Taule, Human vs. machine perceptions on immigration stereotypes, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 8453–8463. URL: https://aclanthology.org/2024.lrec-main.741/.

[31] A. T. Cignarella, M. Sanguinetti, S. Frenda, A. Marra, C. Bosco, V. Basile, QUEEREOTYPES: A Multi-Source Italian Corpus of Stereotypes towards LGBTQIA+ Community Members, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 13429–13441. URL: https://aclanthology.org/2024.lrec-main.1176/.

[32] M. Xia, A. Field, Y. Tsvetkov, Demoting racial bias in hate speech detection, in: Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, 2020, pp. 7–14.

[33] S. Frenda, V. Patti, P. Rosso, Killing me softly: Creative and cognitive aspects of implicitness in abusive language online, Natural Language Engineering 29 (2023) 1516–1537. doi:10.1017/S1351324922000316.

# A. Appendix

## A.1. Positionality Statement

**Annotator 1** describes himself as a 37-years-old men. He is a PhD student, a former activist, and a former humanities teacher.

**Annotator 2** self-describes as a 30-year-old white Italian woman with no migratory background. While she has not experienced migration firsthand, she has lived abroad as an expatriate, navigating international mobility from a position of privilege. Her native language is Italian.

**Annotator 5** self-identifies as a 24-year-old woman second-generation migrant. She is a Master's student with a background in political activism, she has volunteered with non-governmental organizations and has written non-specialist articles about Italian second generation of migrants.

## A.2. Prompts

Presented below is the prompt in Italian:

> *Ti viene fornita in input (Input) una frase estratta dai social media, insieme a tre possibili stereotipi (Opzioni). Il tuo compito è individuare quale stereotipo è implicito nella frase, scegliendo tra le opzioni fornite. Restituisci in output (Output) una singola opzione, sotto forma di lista Python (es. ['Opzione 1']).*
>
> *Input: Mattinata di ieri passata in un'aula di tribunale. 23 udienze al ruolo. Imputati stranieri 19; imputati italiani 4. Imputati stranieri presenti 0; imputati italiani presenti 4. Conclusione (del tutto personale); vengono tutti a delinquere qua e se ne fregano della nostra giustizia*
>
> *Opzioni: [Sono subdoli, Sono pericolosi, Sono una minaccia]*
>
> *Output:*

And its English translation:

> *You are given as input (Input) a sentence extracted from social media, along with three possible stereotypes (Options). Your task is to identify which stereotype is implied in the sentence by selecting one of the provided options. Return as output (Output) a single option in the form of a Python list (e.g., ['Option 1']).*
>
> *Input: Yesterday morning spent in a courtroom. 23 judicial hearings to the register. Foreign defendants 19; Italian defendants 4. Foreign defendants present 0; Italian defendants present 4. Conclusion (entirely personal); they all come here to commit a crime and don't give a damn about our justice*
>
> *Options:[Subtle, Dangerous, Threat]*
>
> *Output:*