An Overview of Current Practices and Recommendations for Working with Stereotypes in NLP

Alessandra Teresa Cignarella[♡] **and Matteo Pellegrini**[♠]
♡ Language and Translation Technology Team (LT3), Ghent University, Belgium
♠ Surrey Morphology Group (SMG), University of Surrey, United Kingdom
alessandrateresa.cignarella@ugent.be, matteo.pellegrini@surrey.ac.uk

Abstract

This article presents a discussion on the main challenges and considerations involved in addressing stereotypes within Natural Language Processing (NLP), and proposes a set of guidelines and recommendations for their treatment in research and resource development. On the one hand, the growing interest in fairness, bias mitigation, and inclusivity has led to an increasing number of studies and datasets dealing with stereotypes; on the other hand, their conceptualization and operationalization remain highly heterogeneous across works. The aim of this article is therefore twofold: (1) to provide a concise yet comprehensive overview of existing annotation schemes highlighting their key features and offering a comparative analysis and (2) to propose a set of tentative guidelines and recommendations to foster clarity when working with stereotypes in NLP. Furthermore, as a case study, we conduct an annotation exercise of a subset of texts from the QUEEREOTYPES dataset, containing stereotypes targeting LGBTQIA+ people, using all labels proposed in prior work to assess their clarity, overlap, and practical usefulness.

Keywords: stereotypes, annotation, agreement, pilot, Italian, LGBTQIA+

Trigger warning: *This paper contains examples of stereotypical and potentially triggering content.*

1. Introduction and Motivation

Stereotypes are exaggerated beliefs associated with a social category (Allport, 1954), they are pervasive social constructs that influence how individuals and groups are perceived, evaluated, and represented (Fiske, 1998). As well as prejudice and discrimination, they can shape social expectations and, in their most extreme forms, contribute to hatred or acts of violence. After having been extensively studied for decades in sociology and psychology, stereotypes have more recently become a central focus of research on fairness, inclusivity, and social bias in NLP, where understanding their linguistic manifestations is crucial for building more equitable and socially aware language technologies (Blodgett et al., 2021).

Yet, despite the growing number of datasets and models addressing stereotypical content (Nangia et al., 2020; Nadeem et al., 2021; Davani et al., 2023; Cignarella et al., 2024; Schmeisser-Nieto et al., 2024, among others), a fundamental challenge remains: the perception of stereotypes is inherently subjective, contextual, and heavily dependent on individual perspectives. What is considered a stereotype, who it targets, who it affects and whether it is harmful or not often depends on cultural backgrounds, lived experiences, linguistic communities, and ideological positions. Treating stereotypes as if they reflected a single, unified ground truth risks flattening the diversity in which they are experienced and interpreted. To adequately capture their complexity there is a growing need for detailed categories and explicitly intersectional approaches (Ma et al., 2023).

The emergence of data perspectivism offers a promising paradigm for confronting this challenge (Cabitza et al., 2023). Perspectivist approaches embrace annotator disagreement rather than eliminating it, and leverage human label variation as a source of information rather than noise (Uma et al., 2021; Leonardelli et al., 2025). Work at the intersection of perspectivism, participatory design, and fairness increasingly demonstrates the value of acknowledging user diversity in annotation and evaluation (Prabhakaran et al., 2021). Nonetheless, existing work on stereotypes in the field of NLP has not yet fully embraced systematic methodologies that explicitly account for a perspectivist approach (aside from some exceptions, i.e., Fraser et al. (2024); Lo et al. (2025)).

As noted by Röttger et al. (2022), "labelled data is the foundation of most natural language processing tasks. However, labelling data is difficult, and there often are diverse valid beliefs about what the correct data labels should be". We fully agree with their observation that "dataset creators should consider the role of annotator subjectivity in the annotation process and either explicitly encourage it or discourage it", and we draw direct inspiration from their approach.

In current NLP research on stereotypes, a wide range of annotation schemes has emerged, each

Proceedings of NLPerspectives @ LREC 2026, pages 112–123

grounded in different theoretical assumptions and targeting different aspects of the phenomenon (Cignarella et al., 2025). These schemes may vary substantially in scope: some focus on stereotypes directed at a single social group, while others consider multiple groups or even multiple dimensions at once, such as the intersection of gender and occupation. In certain cases, stereotypes are annotated alongside related phenomena like polarity, stance or hate speech, whereas other studies treat them in isolation. This heterogeneity makes it difficult to compare resources or understand how specific categories should be interpreted and used. There is therefore a need for greater clarity regarding the utility, granularity, and underlying assumptions of existing annotation categories.

In this article, we review the main annotation labels used so far, discuss their differences and commonalities, and offer commentary on how they can be meaningfully applied within the broader landscape of stereotyping research in NLP.

This article has three main contributions.

- **First.** We offer a critical overview of existing annotation schemes for stereotypes, outlining their defining characteristics and providing a comparative analysis that highlights key points of convergence and divergence across approaches.

- **Second.** To ground this discussion, we conduct a pilot annotation study in which we annotate texts ($N = 100$) from a dataset containing stereotypes targeting the LGBTQIA+ community using the labels and categories drawn from previous studies. The purpose of this annotation exercise is not to introduce a new resource, but rather to use it as a lens through which to examine the practical utility of different annotation categories: which labels are meaningful, which appear redundant or ambiguous, how certain phenomena should (or should not) be annotated, and which dimensions may prove the most helpful for future work in this area.

- **Third.** Finally, on the basis of the outcome of the pilot annotation and the discussion between annotators, we annotate a larger set of social media texts extracted from the same dataset ($N = 400$, see §3 for more details), with an annotation scheme that is a synthesis of what discussed in the previous point. It reuses some of the categories proposed in earlier studies and refines some new ones, complying with the recommendations and best practices identified in the previous stage.

This study is thus intended as an occasion for refining our recommendations, illustrating the challenges and opportunities that arise when operationalizing in real social media data.

2. Related Work

When considering stereotypes in NLP, and especially previous work devoted to data creation and annotation, it can be observed how previous related studies have evolved through three main waves, each characterized by a distinct methodology.

Old-School, the pioneers of sentiment and affect in NLP. Early work (2016–2019) primarily examined harmful language phenomena such as hate speech, offensiveness, and aggressive or stereotyped content. This period was marked by a strong emphasis on linguistically grounded annotation schemes and fine-grained distinctions between related-phenomena. A pioneering example is the Italian Twitter corpus by Sanguinetti et al. (2018), where multi-layer annotations (hate speech, aggressiveness, offensiveness irony) highlighted the pragmatic nuances shaping abusive language online. Parallel efforts on sentiment and polarity, such as SENTIPOLC (Barbieri et al., 2016), integrated sentiment analysis with irony detection, reflecting how pragmatic markers can shift meaning. Additionally, well-known shared tasks such as WASSA[1] advanced research on sentiment, emotion intensity and fine-grained affective states, reinforcing the attention to nuanced expressions beyond simple polarity labels. At the same time, stance detection was explored extensively by Mohammad et al. (2016a,b), who established frameworks that clearly separated stance from sentiment.

Millenials, the psychology-informed and perspective-based generation. A second, more recent wave (2019-2023/2024) began grounding stereotype-related annotation in cognitive and social psychological theory (Fraser et al., 2021). This includes work targeting social groups more explicitly (Nozza et al., 2021), and combining stereotype annotations with hate speech, aggressiveness, offensiveness, irony, sarcasm and stance (Cignarella et al., 2024). Additional contributions in this line include analyses of target-specific slurs (Draetta et al., 2024) and the annotation of forms of discredit to capture subtle mechanisms of stereotyping and prejudice beyond overt hate speech (Bosco et al., 2023; Schmeisser-Nieto et al., 2024).

Gen Z, the free style, identity-aware and perspective-aware approaches. The third, and most recent, line of research (2023/2024-today) explores cognitively aligned representations of stereotype expression, for instance framing stereotypes

[1] https://workshop-wassa.github.io/

as generics by linking a social group to a quality using the reasoning scheme GROUP + relation + QUALITY (Mun et al., 2023) or including free-text descriptions in the form of Subject + Verb + Object or Subject + Noun Phrase patterns (Lo et al., 2025).

The literature on bias and stereotypes in NLP is quite extensive, and the brief overview we proposed is focusing specifically on works that introduce new annotated datasets with clearly defined stereotype-related dimensions. For a broader and more comprehensive perspective on stereotypes and bias, we refer the reader to recent surveys on the topic (Cignarella et al., 2025; Bartl et al., 2025).

3. A Pilot Annotation Study

To better understand how existing stereotype annotation schemes function in practice, we conducted a small-scale pilot study using a subset of texts. This pilot serves as the empirical backbone of our analysis: rather than introducing a new resource, our objective is to test and compare the categories proposed in prior work, assessing their usefulness and limitations when applied to real data.

Specifically, we conduct a focused re-annotation of a subset of QUEEREOTYPES (Cignarella et al., 2024), an Italian social media dataset of stereotypes toward LGBTQIA+ people. It consists of two distinct components: approximately half of the data comes from X and the other half from Facebook. The first portion contains individual tweets annotated for stance and stereotypes, while the latter is organized into status–comment pairs and is annotated for hate, aggressiveness, offensiveness, stereotypes, and irony following the scheme proposed by Sanguinetti et al. (2018). For the purposes of our study, we first harmonized these two sections by completing the annotations for categories that were present in one portion of the dataset but missing in the other.

Then, we only selected the texts containing a stereotype and we enriched them with additional labels drawn from previous research, including: *sentiment and polarity* (Pang and Lee, 2008; Barbieri et al., 2016), *stance* (Mohammad et al., 2016a; Küçük and Can, 2020), *target* (Basile et al., 2019; Nozza et al., 2023), *reported speech* (Schmeisser-Nieto et al., 2022), *slur reclamation* (Kurrek et al., 2020; Draetta et al., 2024), *forms of discredit* (Bosco et al., 2023; Bourgeade et al., 2023), and a *free-text field* (Sap et al., 2020; Lo et al., 2025).

In what follows, we provide details on each of the categories and associated values that we considered when harmonizing and extending the annotations across the two portions of the dataset.

Sentiment/Polarity. A three-way polarity label capturing the affective orientation expressed in the text: *positive*, *neutral*, or *negative*.

Hate Speech. A binary variable (*yes/no*) indicating whether the text contains hateful, hostile, or dehumanizing language directed at an individual or group.

Aggressiveness. A three-level scale assessing the intensity of aggressive expressions: *absent*, *weak*, or *strong*.

Offensiveness. A parallel three-level scale (*absent/weak/strong*) evaluating the degree of insulting, derogatory, or socially inappropriate language.

Irony. A binary label (*yes/no*) indicating the presence of ironic, sarcastic, or otherwise non-literal humorous language.

Stereotype. A binary variable (*yes/no*) marking whether the text conveys a stereotype, i.e., a generalized and often biased attribution of traits or behaviors to a social group.

Stance. A three-way classification capturing the author's position toward the target: *favour*, *neutral*, or *against*.

Target. A coarse-grained label specifying whether the text concerns a *queer*, *non-queer*, or *other* (unspecified/alternative) social group.

Target Specific. An optional free-text field allowing annotators to specify the target at a finer level of granularity (e.g., "gay men", "trans women", "allies").

Reported Speech. A binary label (*yes/no*) marking the presence of quoted or otherwise explicitly reported discourse.

Slur Reclamation. A category identifying the contextual function of slur-related expressions: *slur*, *reclaimed*, or *n/a* (Draetta et al., 2024).

Forms of Discredit. A six-way classification describing how different types of stereotypes undermine or delegitimize the target: *benevolence*, *competence*, *dominance up*, *dominance down*, *affective competence*, and *physical* (Bosco et al., 2023).

Free-text Field. An open field used to encode the stereotype in a schematic proposition, typically following structures such as $S + V + O$ (subject–verb–object) or $S + NP$ (subject–noun phrase) (Lo et al., 2025).

Notes and Comments. A free-text field for annotators to record uncertainties, contextual information, or justifications for labeling decisions.

The two authors of this paper have performed a full and independent annotation of 100 texts (50 tweets and 50 post-comment pairs). Consequently they met to discuss the annotation choices, the disagreements and the difficult cases. We do not report the values of inter-annotator agreement (IAA) for the motivations discussed in in the Limitations section.

The goal of this targeted exercise was not to revise or replace the existing resource, but to explore the complexity of the task and assess clearer

insights into how stereotypes are expressed. By annotating only the stereotype-positive instances with a more detailed taxonomy and a set of labels side-by-side, we were able to interpret the usefulness, clarity, and limitations of each proposed category and to identify which ones meaningfully contribute to a better understanding of stereotyping in NLP.

4. Discussion of Results

From our pilot annotation study, two broad categories of challenges emerge: first, those concerning aspects related to *phenomena, labels and categories* themselves (§ 4.1); and second, those related to *annotation practices and layout*, including data presentation, including methodological choices and features of the platform used (§ 4.2).

4.1. Phenomena, Labels and Categories

4.1.1. Sentiment/Polarity

CHALLENGE: Sentiment annotation proved to be especially problematic. A first issue is conceptual: it is unclear whether annotators should assign sentiment based on an overall intuitive impression or whether they should instead compute something closer to an *algebraic sum* of the positive and negative valence of individual words. These two approaches can produce markedly different labels: if sentiment is treated as a word-level aggregation problem, then annotation becomes almost redundant, since automatic methods (including lexicon-based ones) can compute polarity and even highlight which words contribute to the final score. However, this raises further concerns, as lexicon-based sentiment detection might inherit the biases of the lexicons. The presence of ambivalent sentiment (positive, negative, neutral, or even mixed) adds another layer of complexity, challenging the usefulness of including sentiment as a stand-alone annotation category.

OUR PROPOSAL: Use sentiment annotation only when affective meaning is explicit and directly contributes to the interpretation of the stereotype. In other cases, rely on automatic polarity detection methods as a first pass and let annotators perform only a subsequent light verification step rather than full manual annotation, reducing redundant workload while maintaining quality.

4.1.2. Hate Speech, Aggressiveness and Offensiveness

CHALLENGE: Hate speech, aggressiveness, and offensiveness are intrinsically subjective phenomena. Even with improved guidelines and refined annotation schemes, annotators will inevitably bring their own perspective, background knowledge, and sensitivity to the task. Disagreement can be reduced, but probably never fully avoided, because judgments depend on individual perceptions of what constitutes harm or denigration.

OUR PROPOSAL: Dataset creators should explicitly state whether the goal is to produce a consensus-based gold standard or a subjectivity-aware resource that captures multiple interpretations. Following Röttger et al. (2022), we encourage making annotator subjectivity an intentional design choice: either constrain it (in the case of gold-standard labels) or embrace it (when modelling perspectives).

4.1.3. Irony and Sarcasm

CHALLENGE: Irony and sarcasm typically flip the literal meaning of statements: an utterance may reproduce a stereotype only to mock or reject it. This raises the question of whether the text should be considered as containing a stereotype or not.

Example. *"Why don't you ask your beloved African illegal immigrants what they think of Gay Pride? They are tolerant of homosexuals in their progressive countries; as we know, they adore them"*

This example relies on two classic ironic devices: a rhetorical question in the first part and a false assertion in the second part (Karoui et al., 2017). The ironic meaning becomes interpretable only through world knowledge and pragmatic inference: readers must recognize that the sentence is implausible and therefore intended as its opposite. This makes the utterance a case in which the literal form and the intended meaning diverge, illustrating why irony complicates stereotype annotation and requires explicit annotation of non-literal intent.

OUR PROPOSAL: Annotate the presence of the stereotypical content and add a separate flag for the presence of irony or sarcasm. This keeps literal content and intended meaning distinct and avoids mislabeling anti-stereotypical discourse.

4.1.4. Explicit and Implicit Stereotypes

CHALLENGE: Identifying whether a stereotype is present in a text is far from straightforward. Firstly, the field lacks a clear, operationalized definition of what constitutes a stereotype in NLP (Devinney, 2025). As a consequence, annotators may rely on personal intuition rather than shared criteria. Secondly, as noted by Schmeisser-Nieto et al. (2022), many stereotypical meanings are not stated overtly but must be inferred from background knowledge, pragmatic cues, or cultural assumptions.

Example. *"@user MY AUNT MARIA SAYS YOU'RE ALSO HOMOSEXUAL, I DON'T BELIEVE IT, YOU'RE SUCH A HANDSOME MAN, OR NOT?"*

This text suggests that being homosexual is incompatible with being handsome. The stereotype can thus be detected only via an inferential step.

OUR PROPOSAL: We propose to clearly distinguish between *explicit* and *implicit* stereotypes. Explicit stereotypes are those directly stated in the text (e.g., "LGBT people are not pure"), for which higher agreement is expected. Implicit stereotypes, by contrast, rely on presuppositions, implicatures or culturally-shared associations. Because different readers may draw different inferences, especially for implicit cases it is crucial to collect multiple voices rather than collapsing perspectives into a single viewpoint.

4.1.5. Identification of Specific Targets

CHALLENGE: Annotating the target of a stereotype is not always straightforward. While a coarse-grained label such as *queer*, *non-queer* or *other* is useful for ensuring consistency. However, in some utterances, the target of the stereotype is not overtly specified, but there is only a vague reference, e.g., to a generic "they". In other cases, the utterance concerns a specific subgroup whose identity cannot be captured adequately by a single coarse category.

OUR PROPOSAL: We adopt a two-level annotation strategy. First, annotators assign a coarse-grained **Target** label (*queer*, *non-queer*, *other*). Second, when the target can be identified more precisely, annotators may use the optional **Target (Specific)** free-text field to capture finer-grained subgroups (e.g., "gay men", "trans women", "allies"). This approach allows us to maintain a unified structure while preserving valuable detail when available.

4.1.6. Direct and Indirect Target

CHALLENGE: Stereotypical utterances typically have the underlying structure "Target group X has characteristic Y". In this paper, we focus on stereotypes involving queer individuals, who may appear either as the *direct* or the *indirect* target of the stereotype. We therefore distinguish between: (i) cases in which queer individuals are the direct target X of the stereotype, corresponding to statements of the form "All queer individuals have characteristic Y", and (iia-b) cases in which another individual or group is the direct target X, and the stereotypical association involves queer identities or queer-related evaluations only indirectly.

The latter category includes two common patterns: (iia) utterances in which a non-queer target is stereotypically associated with queerness as characteristic Y, and (iib) utterances in which the target group is described negatively because they support queer individuals.

Examples.

(i) Nowadays we find gays and transsexuals everywhere because they are included in all the TV programs.

(iia) @user If that's really the case, there's also the aggravating circumstance of homophobia... considering the not exactly masculine reaction of the violent robber

(iib) They must be fake priests, surely left-leaning, probably atheists too. [Referring to priests who take part in marches against homophobia]

OUR PROPOSAL: All of the above configurations should be treated as relevant for the study of stereotypes involving the LGBTQIA+ community. However, we argue that it is important to encode the distinction between them in the annotation scheme, as this enables more fine-grained analyses of how queer identities are targeted in discourse. To this end, we introduce a dedicated annotation category, **directness**, with three possible values:

- **(i) DIRECT**: queer individuals are the direct target of the stereotype.
- **(iia) INDIRECT-CHARACTERISTIC**: another individual or group is the target, but they are stereotypically associated with queerness.
- **(iib) INDIRECT-ALLIES**: another individual or group is targeted negatively because they support queer individuals.

We could also distinguish cases where queer individuals appear as initiators or as recipients of the action, an idea reminiscent of semantic-role labeling (agent/patient). This point connects naturally to the free-text rewriting of stereotypes (Lo et al., 2025) and will be elaborated further in Section 4.1.11.

4.1.7. Prejudice and Discredit

CHALLENGE: In some cases, even when a stereotype is clearly present in the text, it does not necessarily express discredit toward queer individuals. This situation may arise for several reasons: (a) the stereotype attributes to queer individuals a characteristic that is not negative *per se*; (b) the stereotype is generic or it is not-clear which feature is being attributed to queer individuals; or (c) queer individuals are not the direct target of the stereotype but appear only indirectly (see § 4.1.6).

Examples.

(a) "Eleonora, men are all the same" as my mother says. I think I might want to become a lesbian at this point.

(b) @user, what a terrible combination you are... Democratic Party supporter, lawyer, AC Milan fan... You're just missing being a lesbian and you'd be all set...

(c) But those people in show business... singers, actors, fashion designers... are they all gays or lesbians? What kind of atmosphere is there over there?

OUR PROPOSAL: We recommend distinguishing between *derogatory* and *non-derogatory* stereotypes.[2] Based on this distinction, we propose annotating discredit only when the stereotype is classified as **derogatory** (see § 4.1.8). This makes

[2]Here we intentionally avoid the terms "positive" and "negative" since even non-derogatory stereotypes, such

the discredit label an optional, downstream category applied in a cascade fashion: first determine whether a stereotype is present, then whether it is derogatory, and only if it is annotate discredit.

4.1.8. Forms of Discredit

CHALLENGE: Existing discredit categories were originally designed for other target groups such as migrants (Bosco et al., 2023; Bourgeade et al., 2023) and may not map neatly to the LGBTQIA+ community as a target.

Furthermore, the presence of discredit presupposes that the underlying stereotype is derogatory (see § 4.1.7): if the stereotype is non-derogatory, then discredit is absent and thus conceptually inappropriate.

OUR PROPOSAL: We recommend revisiting discredit categories and adapting them specifically for queer-related contexts rather than importing labels designed for other targets. This involves (i) ensuring that discredit is annotated only when the stereotype is classified as *derogatory*; (ii) clarifying the terminology, avoiding fuzzy or overlapping labels and re-evaluating whether certain distinctions are meaningful or empirically grounded for LGBTQIA+ stereotypes.

Example.
A stereotype category (PHYSICAL) developed for migrants typically refers to dirtiness, diseases, or the idea that migrants "carry illnesses" (Bosco et al., 2023; Schmeisser-Nieto et al., 2024). For queer-related stereotypes, however, this dimension needs to be rethought. In queer contexts, physicality often appears through markers such as clothing style, haircut, grooming, or body modifications treated as stereotypical "signals" of queerness. Some cases still invoke health-related stigma (e.g., *"National data showed infections among homosexual men... why not ask why?"*), while others rely on appearance-based cues (e.g., *"Typical lesbian haircut and short nails"*). This suggests expanding the physical category to include both health-based stigmas and appearance-related markers commonly used to stereotype queer individuals.

4.1.9. Stance, Reported Speech and Counterspeech

CHALLENGE: Some posts or tweets reproduce a stereotype not to endorse it, but to comment on it or explicitly reject it. In such cases, the stereotypical content appears only within reported speech, while the author's own stance is oppositional. This creates a misalignment between the *literal content* (which may contain harmful or stereotypical expressions) and the *author's intention* (which may be supportive of queer individuals). These cases also raise questions about the annotation of stance: if someone cites a derogatory stereotype against the LGBTQIA+ community to reject it, the stance of the stereotype towards the community will be trivially negative, but the stance of the author will be positive. As noted by Schmeisser-Nieto et al. (2022), reported speech, counterspeech, and euphemisms are rare and therefore often grouped together, yet they are pragmatically distinct and particularly relevant for tasks such as automatic moderation, where it is crucial to differentiate authors who propagate stereotypes from those who criticize them.

Example.
"I think you are a lesbian." Why? Because I wear jeans and T-shirts and do not use make-up and do not care about womanly things? WAKE UP, I AM STRAIGHT AND LESBIANS CAN BE MODELS WEARING GUCCI FROM HEAD TO TOE. Stop stereotyping.

OUR PROPOSAL: We propose treating reported stereotypes as regular stereotype occurrences but marking them explicitly with a dedicated **"Reported Speech"** flag. Annotators should (i) identify the presence of a stereotype in the reported content, and (ii) annotate the author's stance and not the stance conveyed by the stereotypical utterance. This allows for cases where a neutral or negative sentiment from a reported speech co-occurs with an author's stance supportive of queer individuals.

4.1.10. Slur Reclamation

CHALLENGE: Slurs can appear either as clear insults or as reclaimed, in-group expressions (Ferrando et al., 2026). The same term may therefore be derogatory in one context and identity-affirming or playful in another. Because reclaiming depends on speaker identity, audience, and context, treating all occurrences uniformly as harmful would conflate hostile uses with in-group language practices.

OUR PROPOSAL: We adopt a simple three-way label for the slur category. We use *slur* only when the term is clearly derogatory; *reclaimed* for in-group uses, and *none* where there is no slur (although we may still see offensiveness or aggressiveness which could be annotated as per § 4.1.2).

4.1.11. Free-Text

CHALLENGE: Free-text fields enable annotators to capture stereotype constructions beyond predefined labels, including generic GROUP + RELATION + QUALITY statements (Mun et al., 2023) or syntactically grounded S + V + O sentences (Lo et al., 2025). In addition, free-text annotation can be used to record entailments and implicit meanings (Sap et al., 2020). However, this flexibility results in highly variable outputs: annotators differ in wording, level of detail and linguistic focus. Such variability may complicate aggregation and can blur the

as those assigning supposedly positive traits, can reproduce harmful biases and reinforce structural inequalities (e.g., "women are naturally better at caregiving").

line between what is stated in the text and what is inferred by the annotator.

OUR PROPOSAL: We recommend emphasising that free text annotation are not to be intended as completely unconstrained and descriptions, but have to conform with the structure of the S + V + O or GROUP + RELATION + QUALITY prompt. This avoids obtaining noisy results, that might be too difficult to aggregate or compare and facilitates automatic extraction (Felkner et al., 2023; Lo et al., 2025).

4.2. Annotation Practices and Layout

4.2.1. Annotation Platform

CHALLENGE: Annotation platforms differ widely in functionality and usability: simple layouts, such as drop-down menus on spreadsheets, offer flexibility but lack interface support; tools like *LabelStudio* provide richer workflows but may require technical setup; crowdsourcing platforms (e.g., *Prolific*) introduce additional variability in annotator background and quality control; proprietary solutions can limit transparency and reproducibility.

OUR PROPOSAL: We recommend selecting the right medium based on task complexity and the need for controlled guidance. For tasks involving nuanced or implicit stereotypes, interfaces that support clear instructions, validation rules, and structured free-text fields are preferable. Regardless of the platform, we encourage documenting interface design choices in a Data Statement (Bender and Friedman, 2018; McMillan-Major et al., 2024), exporting data in interoperable formats and conducting small pilot rounds to ensure that the tool supports reliable and consistent annotation.

4.2.2. Annotation Granularity

CHALLENGE: Stereotypes may appear at different textual levels: entire documents, social media threads, single posts or paragraphs, or specific spans of text. If the intended unit of annotation is not clearly defined, annotators may rely on different amounts of context, leading to inconsistent judgments and reduced reliability.

Currently, the annotation is performed for each facebook post or tweet, but in some cases it can be useful or necessary to identify shorter spans of texts to be annotated. In general, this can be useful when the text of the post/tweet is very long, and in particular if more than one stereotype is present, possibly of different types. In the following example, stereotypes against both women and gay people can be identified in different parts of the text.

Example.

If you think that those who are contesting the manifesto are all women, it gives me shivers. They are those who want abortion to murder defenseless beings by right, but ask for the adoptability of other people's children and the uterus of others for rent to give children to GAY people. What world are we going to where human life can be suppressed by law and it is believed to be a right to be able to suppress it.. Women....shame on you and you also have the pretense of calling yourselves mothers??

OUR PROPOSAL: We propose explicitly stating the annotation level and providing minimal examples to illustrate it. If annotation below the level of the post/tweet is deemed necessary, it could be achieved in different ways, for instance by performing an *a priori* segmentation of longer texts into phrases/sentences/conversation turns (depending on the task to be accomplished by the researcher) or keeping the whole textual unity together despite its length but identifying and annotating shorter spans of text, using tools such as *LabelStudio* or similar. The choice between these two options depends on the specific needs of different studies.

4.2.3. Order of Annotations

CHALLENGE: The order in which different annotations is performed matters (Beck et al., 2024). In some cases, this is trivial (e.g., the category "target specific" cannot but be annotated after the category "target"), but in other cases the reason can be subtler (e.g., the stance of a text should be more precisely interpreted as the stance of the writer about the target of the stereotype, so it should be annotated after having identified the target of a stereotype).

Another example is the free-text description of the stereotype: if it is performed after other annotations, it can be influenced by the labels proposed as options of the other annotations. For instance, the values of the category pertaining to the type of discredit can be suggestive of the content of the stereotype, and the identification of the target might determine the template structure.

OUR PROPOSAL: The order in which annotations are presented should be considered carefully and instructions should be given to annotators on the order in which they should be performed. For instance, attention should be paid to whether the annotation is performed "horizontally" (i.e., all categories are annotated for a single text before going on to the next text) or "vertically" (i.e., all texts are annotated for a given category before going on to the next category). If there are logical interdependencies like the ones we just discussed, the annotation needs to be performed horizontally, but other categories can be more efficiently and consistently annotated vertically.

4.2.4. Availability of Context

CHALLENGE: In some cases, depending on how data was retrieved, contextual information might be missing, so that it is impossible to provide a fine-grained annotation on some aspects, or at least

it is necessary for annotators to draw non-trivial inferences. For instance, the Facebook data of our sample consists of a set of texts, each consisting of status-comment pair. However, in some cases the text of the original status make crucial reference to a picture, or to a web page, that were not available at the time of re-annotation. Consequently, sometimes non-trivial inferences are needed to understand the nature and details of the stereotype.

In the example below, commenting on a picture referring to male same-sex parenting, regarding comment (i) it can be inferred that the stereotype has something to do with gay males being unfit for parenthood, but for comment (ii) it is difficult to provide a fine grained annotation of all the categories involved without further information on the context.

Examples.

STATUS: *SOMETIMES A PHOTOGRAPH IS WORTH A THOUSAND WORDS*

COMMENT: *(i) What violence to this poor little one. He wants mommy and he wants to suck milk.*

COMMENT: *(ii) Poor baby*

The most common scenario when dealing with social media data is the presence URLs that may point to images, videos, or external content crucial for interpreting the message. Allowing annotators to open links introduces multimodal information that may be necessary for understanding stereotypes, but it also creates inconsistencies: different annotators may access different versions of the linked content, encounter unavailable pages, or rely on information that is not preserved in the dataset. This raises concerns about reproducibility and comparability, both among annotators and between human annotations and downstream systems, which may not have access to the same external resources.

OUR PROPOSAL: We recommend defining a clear policy on how to handle external context, aligned with the dataset's goals. If external content is essential, linked material should be archived or embedded (e.g., screenshots, textual extracts) and made available also offline, so that all annotators and systems access the same information. If multimodal access is not feasible or not in scope, annotators should be instructed *not* to open external context, and guidelines should clarify that the annotation must rely solely on the text provided.

5. Open Problems and Discussion

We highlight a set of open issues that, rather than calling for immediate solutions, we offer as prompts for discussion at the workshop.

- **New layers on old data.** Adding new annotation layers to existing datasets creates dependencies between past and present labels. Prior annotations can implicitly bias new judgments, raising concerns about reinterpretation of legacy data and the long-term consistency of datasets that accumulate layers over time.
- **Inter-annotator subjectivity across dependent layers.** When new layers depend on earlier interpretive decisions, disagreements become harder to resolve. Annotators may not perceive a stereotype where a previous annotator did, yet the earlier label remains part of the data. These inconsistencies highlight the fragility of sequential and interdependent annotations.
- **Annotator sensitivity and topic familiarity.** Annotators vary in how attuned they are to sensitive or domain-specific content. Limited familiarity, personal distance from the topic, or fatigue from prolonged exposure can all affect perception and judgment (Beck et al., 2024).
- **Annotator self-consistency.** Even the same annotator may label the same content differently when revisiting it later. Such intra-annotator variation raises questions about the stability of stereotype-related judgments and how to interpret labels that reflect inherently fluid perceptions.

We present these issues not as problems to be resolved here, but as starting points for discussion on how to navigate the complexities of stereotype-related annotation.

6. Conclusion

In this paper, we summarized current practices, challenges, and emerging considerations for annotating stereotypes in NLP. This empirical survey of existing approaches reveals that stereotype annotation remains a complex task shaped not only by the categories themselves, but also by the nature of the data, the design of annotation schemes, and the practical conditions under which annotators work. Choices regarding granularity, annotation layout, interface design, and the introduction of new layers all have effects on annotator reasoning, consistency and interpretation. Moreover, the inherently subjective and context-dependent nature of stereotypes introduces variability that cannot be fully eliminated, but can be better understood and anticipated.

Rather than prescribing definitive solutions, our aim has been to highlight areas where further reflection and discussion are needed. As research on stereotypes continues to evolve, so too must our annotation frameworks. We hope that the considerations raised here encourage more transparent, context-aware, and reflexive annotation practices and that they support the development of

datasets that more accurately capture the nuanced ways in which stereotypes appear in language.

Data Availability

Due to licensing and usage restrictions, the original QUEEREOTYPES dataset can be released only privately, upon request. Interested parties will be required to reach out to the first author, complete an agreement form, outlining the specifics of their research in order to obtain the password that protects the files. It is essential for them to ensure compliance with GDPR regulations and other policies from both X and Facebook.

Limitations

Our work presents some limitations. First, the analysis is restricted to a single target category and a single language, which limits the generalisability of our reflections to multilingual or cross-cultural contexts. Second, we build on an existing resource, and we do not question the reliability of inherited labels, nor fully control for how previous annotations may influence new layers.

Regarding inter-annotator agreement (IAA), although we computed it, we do not report the results because they are not informative for our setting: (i) several categories were inherited or only partially re-annotated (*Hate Speech*, *Aggressiveness*, *Offensiveness*, *Irony*, *Stance*, with *Stereotype* trivially always $= 1$); (ii) some labels are almost deterministic from the text or scheme (*Target*, *Reported speech*, *Slurs Reappropriation*); (iii) *Forms of discredit* involves overlapping classes with no straightforward agreement metric; and (iv) free-text fields (*Target Specific*, *S+V+O*/*S+NP*) are not suited to standard IAA measures.

Finally, only two annotators with similar backgrounds contributed to the new layers, which limits the diversity of perspectives and increases the risk of subjective bias. For these reasons, we avoid drawing overly definitive conclusions and view our findings as a starting point for further discussion.

Acknowledgements

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. Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA).

Bibliographical References

G. W. Allport. 1954. *The nature of prejudice*, addison-wesley edition. Addison-Wesley.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTIment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016*, volume 1749 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*, 57(6):1–36.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 54–63. Association for Computational Linguistics.

Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. 2024. Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *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)*, pages 1004–1015.

Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP. *Information Processing & Management*, 60(1):103118.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6860–6868. AAAI Press.

Alessandra Teresa Cignarella, Anastasia Giachanou, and Els Lefever. 2025. A survey on stereotype detection in natural language processing. *ACM Computing Surveys*, 58(5).

Alessandra Teresa Cignarella, Manuela Sanguinetti, Simona Frenda, Andrea Marra, Cristina Bosco, and Valerio Basile. 2024. QUEEREOTYPES: A Multi-Source Italian Corpus of Stereotypes towards LGBTQIA+ Community Members. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13429–13441, Torino, Italia. ELRA and ICCL.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Hannah Devinney. 2025. Power(ful) Associations: Rethinking "Stereotype" for NLP. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 52–58. Association for Computational Linguistics.

Lia Draetta, Chiara Ferrando, Marco Cuccarini, Liam James, and Viviana Patti. 2024. ReCLAIM Project: Exploring Italian Slurs Reappropriation with Large Language Models. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140.

Chiara Ferrando, Lia Draetta, Marco Madeddu, Mae Sosto, Viviana Patti, Paolo Rosso, Cristina Bosco, Jacinto Mata, and Estrella Gualda. 2026. MultiPRIDE at EVALITA 2026: Overview of the Multilingual Automatic Detection of Slur Reclamation in the LGBTQ+ Context Task. In *Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, February 26th-27th, 2026*. CEUR-WS.org.

Susan T Fiske. 1998. Stereotyping, prejudice, and discrimination. In *The handbook of social psychology*, pages 357–411. McGraw-Hill.

Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2024. How Does Stereotype Content Differ across Data Sources? In *Proceedings of the 13th joint conference on lexical and computational semantics (* SEM 2024)*, pages 18–34.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *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)*, pages 600–616, Online. Association for Computational Linguistics.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for

Online Slur Usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.

Soda Marem Lo, Marco Antonio Stranisci, Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Elisabetta Ježek, and Viviana Patti. 2025. Subjectivity in stereotypes against migrants in italian: An experimental annotation procedure. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 603–612. CEUR Workshop Proceedings.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023. Intersectional stereotypes in large language models: Dataset and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore. Association for Computational Linguistics.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2024. Data Statements: From Technical Concept to Community Practice. *ACM Journal on Responsible Computing*, 1(1).

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A Dataset for Detecting Stance in Tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 31–41. The Association for Computer Linguistics.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *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)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406.

Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, Tommaso Caselli, and Viviana Patti. 2023. HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023*, volume 3473 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. 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*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490. Association for Computational Linguistics.

Wolfgang Schmeisser-Nieto, Montserrat Nofre, and Mariona Taulé. 2022. Criteria for the annotation of implicit stereotypes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 753–762.

Wolfgang S Schmeisser-Nieto, Alessandra Teresa Cignarella, Tom Bourgeade, Simona Frenda, Alejandro Ariza-Casabona, Mario Laurent, Paolo Giovanni Cicirelli, Andrea Marra, Giuseppe Corbelli, Farah Benamara, et al. 2024. Stereohoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*, pages 1–39.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.