Studying Reactions to Stereotypes in Teenagers: an Annotated Italian Dataset

Elisa Chierchiello*[1], Tom Bourgeade*[1], Giacomo Ricci[1], Cristina Bosco[1] and Francesca D'Errico[2]

[1] Dipartimento di Informatica - Università di Torino, Italy
elisa.chierchiell@edu.unito.it, tom.bourgeade@unito.it
giacomo.ricci@edu.unito.it, cristina.bosco@unito.it
[2] Dipartimento Formazione, Psicologia, Comunicazione - Università di Bari, Italy
francesca.derrico@uniba.it@uniba.it

Abstract

The paper introduces a novel corpus collected in a set of experiments in Italian schools, annotated for the presence of stereotypes, and related categories. It consists of comments written by teenage students in reaction to fabricated fake news, designed to elicit prejudiced responses, by featuring racial stereotypes. We make use of an annotation scheme which takes into account the implicit or explicit nature of different instances of stereotypes, alongside their forms of discredit. We also annotate the stance of the commenter towards the news article, using a schema inspired by rumor and fake news stance detection tasks. Through this rarely studied setting, we provide a preliminary exploration of the production of stereotypes in a more controlled context. Alongside this novel dataset, we provide both quantitative and qualitative analyses of these reactions, to validate the categories used in their annotation. Through this work, we hope to increase the diversity of available data in the study of the propagation and the dynamics of negative stereotypes.

Keywords: Stereotypes, Italian, Annotated Corpus, Linguistic Analysis

1. Introduction

Stereotypes are often used to describe people who belong to a different group, have a different physical appearance or different social behavior. They are useful to reduce the cognitive complexity we have to deal with when we are confronted with different situations. However, negative stereotypes often occur in connection with hate speech and discrimination, phenomena that have become more widespread with the increasing use of social media as platforms for communication and exchange.

This work addresses the study of negative stereotypes from a perspective that encompasses both psychology and computational linguistics. We present a novel corpus in which racial stereotypes are annotated, namely the STERHEOSCHOOL corpus. It consists of a selection of data collected in Italian schools as part of an experiment conducted by a group of social psychologists (Corbelli et al., 2023; D'Errico et al., 2023) within the STERHEOTYPES project[1]. More precisely, this corpus includes two racial hoaxes and the reactions provided by teenagers that read them. The hoaxes are artificially created news articles, presented as if they were recorded via a cell phone interface, and designed to elicit reactions in readers that may contain stereotypes. For each news item, readers were asked to comment on the news in general, as well as the main character of the articles in particular. These comments are moreover associated with metadata, such as age and declared gender of the author, which enable some analyses of the annotated labels' distribution.

We applied to the news and comments provided by the readers an annotation scheme which includes two different main categories and related subcategories, inspired by two annotation schemas applied on other corpora developed as part of the STERHEOTYPES project. The first category concerns the presence of stereotypes as implicitly or explicitly expressed, then the forms of discredit used against targets of these stereotypes in the news items (Bourgeade et al., 2023). The second main category concerns the annotation of the commenters' stance concerning the news items (Cignarella et al., 2023), linked to rumor and fake news stance detection (Küçük and Can, 2020), which are relevant to the context of this dataset of reactions to fabricated fake news articles. In the application of the annotation schema, we addressed some of the challenges related to the specific structure of the data collected by psychologists.

By providing data collected in schools and generated by teenagers, this study aims at filling a gap in the literature. Teenagers are indeed an underrepresented category in data annotated for

*These two first authors contributed equally to the paper.

[1]STERHEOTYPES (Studying European Racial Hoaxes and sterEOTYPES) is an international project funded by Compagnia di San Paolo and VolksWagen Stiftung

TRAC-2024 Workshop, pages 115–125

text classification tasks, since almost all the available corpora are composed of messages drawn from social media platforms (non-frequented by adolescents) and rarely associated with information about the age of the authors. As such, the main contributions of this paper are: (1) we provide a novel annotated resource for the study of racial stereotypes and related categories in Italian; (2) we explore stereotypes in an uncommon setting and genre, i.e. fabricated fake news developed for studying the reactions of teenage students to racial stereotypes; (3) finally, we provide quantitative and qualitative analysis of the annotated data, through the lenses of lexical and linguistic analysis.

The paper is organized as follows: the next section briefly introduces the related work. Section 3 describes the corpus, focusing on the collection and annotation of the data. In Section 4, we provide a quantitative lexical analysis of the annotated data, followed by qualitative linguistic observations in Section 5. Finally, we provide a discussion and some conclusions.

2. Related Work

The notions of stereotype and prejudice are often used almost as synonyms since stereotypes are the cognitive nucleus of prejudice, which assumes, in turn, the face of discrimination, or racist and hateful behaviour in social interactions often identified as Hate Speech (HS).

According to social psychology (Allport, 1954), the stereotype is a firmly held association between a social group and some physical, mental, behavioral features or occupational quality. It is a form of generalization about a group of people, in which the same characteristics are assigned to virtually all members of the group, regardless of the actual and meaningful variation among the group members. The generation of stereotypes is the result of an automatic mental process, i.e. categorization, but their diffusion depends on socialization that very often employs mass media (Vaes et al., 2017; D'Errico and Papapicco, 2022).
Negative stereotypes can often start the development of prejudices about a social group and of specific behavioral attitudes against it in general or some of its members in particular. Prejudice can be in turn expressed through verbal forms of racism or discrimination, in the literature, indicated as discredit (van Dijk, 2016).

Within the context of computational linguistics, in the last few years, stereotypes started to raise some interest, but very limited when compared with the interest devoted to HS and closely related phenomena, such as abusive language and toxicity or misogyny as the rest of this section shows. The identification of HS in its various forms is based on multidisciplinary approaches (like social psychology, law and social sciences), but NLP seems in effect to play an important role in their investigation. Among the several events and shared tasks held about these topics and reflecting the interest in hate speech by the computational linguistics community, we can cite those organized in the international evaluation campaign *SemEval 2019*, *SemEval 2020* and *Semeval2023*: the *Shared Task 5 on Hate Speech Detection against Immigrants and Women* for English and Spanish (Basile et al., 2019)[2], the task 6 of SemEval 2019 on *Identifying and Categorizing Offensive Language in Social Media* (OffensEval)[3] (Zampieri et al., 2019), *OffensEval 2: Multilingual Offensive Language Identification in Social Media* (Zampieri et al., 2020)[4] and *Task 10: Towards Explainable Detection of Online Sexism* (Kirk et al., 2023). Another relevant event is the Workshop on Online Abuse and Harms (WOAH) whose first edition was organized in 2017 and the last in 2023 (Chung et al., 2023).
For Italian, a task about HS has been proposed for the first time in *Evalita 2018*, i.e. *Hate Speech Detection* (HaSpeeDe) held in 2018 (Bosco et al., 2018) and then in the two following editions of this campaign in 2020 and 2023 respectively[5] (Lai et al., 2023; Sanguinetti et al., 2020) in which hateful contents about different targets have been analyzed.
Other related events are the tracks on *Automatic Misogyny Identification* (AMI) (Fersini et al., 2018b) and on *Authorship and aggressiveness analysis* (MEX-A3T) (Carmona et al., 2018) proposed at the 2018 edition of *IberEval*, the *Automatic Misogyny Identification* task at *Evalita 2018* (Fersini et al., 2018a). For Spanish other evaluation exercises were organized recently such as *DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish* (Alejandro Ariza-Casabona, 2022) and *NewsCom-TOX: a corpus of comments on news articles annotated for toxicity in Spanish* (Mariona Taulé, 2023).

These tasks were highly participated and this indicates the interest of the community towards HS and encouraged the proposal of various editions of these events. Being the techniques used for detecting HS are mainly based on machine learn-

[2] https://competitions.codalab.org/competitions/19935

[3] https://sites.google.com/site/offensevalsharedtask/offenseval2019.

[4] https://sites.google.com/site/offensevalsharedtask/

[5] http://www.di.unito.it/~tutreeb/haspeede-evalita20/ and http://www.di.unito.it/~tutreeb/haspeede-evalita23/

ing, they require annotated corpora. In most cases, the data used for building them are extracted from social media, such as Twitter and FaceBook from where are extracted the data used for the first *HaSpeeDe* task (Bosco et al., 2018).
Nevertheless, while several corpora, used as benchmarks in shared tasks or not, include the annotation of different phenomena related to HS, only very few are also annotated to make explicit the presence of stereotypes. Among them, we can especially cite the dataset exploited in the *Hate Speech Detection* (HaSpeeDe) and HaSpeeDe 2020 (Bosco et al., 2018; Sanguinetti et al., 2020). In this case, only a basic form of annotation is used to make explicit the presence (or absence) of the stereotype. In the dataset developed for the DETEST a finer-grained annotation has been applied which includes also the category of the stereotype target and a mark for implicit (Alejandro Ariza-Casabona, 2022).
Other more recently developed corpora include also or only (without considering HS) the annotation at finer-grained level of stereotype, in particular, the corpora that inspired our annotation scheme and we cited above, i.e. (Bourgeade et al., 2023) and (Cignarella et al., 2023).

The scarce availability of resources annotated for stereotype explains the limited possibility of research activities and development of tools for the automatic detection of this phenomenon, that is considering especially challenging. It can be indeed observed that also in shared tasks providing datasets where they were annotated, systems were not properly tested for their ability to detect this category, with the only exception of the HaSpeeDe shared task organized in 2020 (Sanguinetti et al., 2020) where a pilot subtask was about the detection of stereotypes. Only very recently some work has been issued about stereotype where a computational view is provided (Fraser et al., 2022) and a task related to the detection of stereotype has been devised within PAN: *Profiling Irony and Stereotype Spreaders on Twitter* (IROSTEREO 2022)[6].

Some dimensions can make also more challenging the detection of stereotypes, such as the fact that they can be expressed both in explicit and explicit form. An interesting analysis of this topic is provided in Schmeisser-Nieto et al. (2022), where the implicitness of stereotype is especially observed. Given the scarcity of studies in this regard, it is important to refine the ability to detect racial stereotypes even when they are expressed implicitly. The implicit structure is particularly appropriate for conveying messages that contain stereotypes, as it presents two irrefutable advantages: it lures the listener in while protecting the speaker (Domaneschi and Penco, 2016). As highlighted in (Reboul, 2011), everyone tends to fall victim to an egocentric bias, which leads to preferring one's beliefs to those of others, even when these are generated from an external input, such as a message that we listen to or read. Therefore, the longer the chain of inferences we use to reconstruct the message, the more we tend to accept it without objections or criticism.
It is no coincidence that the distinction between implicitness and explicitness, problematic as it may be, consists of the distinction between saying and implying. In other words, an implicit statement conveys the message intended by the speaker, but it does not match the sentence that is spoken, which is why its detection may be difficult for humans, and even for machines.

3. Dataset

The corpus described in Corbelli et al. (2023) and D'Errico et al. (2023) comprises a curated collection of racial hoaxes relating to people from European and African origins. Each racial hoax in the dataset is uniquely identified by an ID. Accompanying these hoaxes are two sets of commentaries from the students who analyzed them: one about the news (*commento notizia*) and one about the leading actor of the news (*commento protagonista*), which have been merged into a single comment (*commento unico*) in the STERHEOSCHOOL corpus. In addition to the textual content of the hoaxes and the commentaries, the dataset includes demographic annotations from the student participants, specifically their self-reported age and gender.
In the STERHEOSCHOOL dataset, only two distinct racial hoaxes serve as focal points for analysis and discussion. They were selected from the larger corpus cited above because they particularly emphasize the complex interplay of race, media and social perception, but also because they are the ones around which the most commentary revolves. The first hoax (see Figure 1a) involves a fabricated story centered around a group of individuals from Naples, Italy. This narrative was designed to provoke racial biases by depicting the Neapolitan protagonists in a manner that reinforces negative stereotypes, despite the story's complete lack of factual basis. The second hoax (see Figure 1b) shifts the geographical and cultural context to Africa, presenting a concocted incident involving African protagonists. Similar to the first, this hoax was crafted to elicit prejudiced reactions by exploiting and distorting cultural and racial stereotypes associated with Africans.

These two articles were split into 9 variants,

[6] https://pan.bis.de/clef22/pan22-web/author-profiling.html#task-committee

each a combination of 3 different ways of presenting the artificial impact of the article (high number of "likes", low number of "likes", no number) with 3 different types of reactions to the article ("positive" comments, "negative" comments, no comments). In this work we do not exploit this aspect directly, and we consider only the two fabricated news articles and the associated students' comments. In total, after filtering (empty or otherwise not exploitable comments), 1147 student comments were collected and annotated for the two articles (see Subsection 3.3 and Figure 2 for the annotation and distribution of labels).

Both hoaxes were meticulously chosen for their capacity to illuminate the mechanisms through which racial prejudices are constructed and perpetuated in society. Through the lens of these fabricated stories, the dataset captures the reactions of adolescents, offering valuable insights into their perception of race and the influence of media on their understanding of racial dynamics. The comments on these hoaxes, derived from a diverse group of students, reveal a range of perspectives that reflect varying degrees of awareness, bias, and critical thinking regarding race and media representation. By examining these two contrasting yet similarly intentioned hoaxes, the dataset provides a unique opportunity to explore and address the challenges of racial misinformation and its impact on young minds in different cultural contexts.

This dataset is useful to facilitate a comprehensive analysis of the impact of racial hoaxes on adolescent perceptions and to foster a deeper understanding of racial issues among young people.

3.1. Collection

As far as the collection of the data, the research was split into two phases; in the initial phase, conducted using computers in the school's labs, the participants filled out a preliminary set of tests and surveys. This was done to gather fundamental socio-demographic data and to evaluate affective prejudice, both active and inhibitory self-regulatory efficacy, as well as implicit biases. In the subsequent phase, which took place a week later, the same group of students were introduced to a novel analytical tool that was both quantitative and qualitative in nature, created via Google Forms anonymously.

Through establishing a fictional scenario where the student plays a role in an online newsroom, a deliberate effort was made to help the participant identify the communicative and substantial elements that define a racial hoax. This process also involved evaluating their capability to learn and identify racially motivated misinformation. Subsequently, the adolescents were asked to reinterpret the same news piece from the perspective of the immigrant involved in the story. This exercise aimed to encourage them to merge two narratives: the initial misleading one and the second one centered on the immigrant's viewpoint. After this activity, the students' inclination to rationalize ethnic-based moral transgressions (termed as Ethnic Moral Disengagement) was reassessed using the same criteria as in the first stage. The total time required to complete the entire exercise ranged between 30 and 50 minutes.

3.2. Annotation

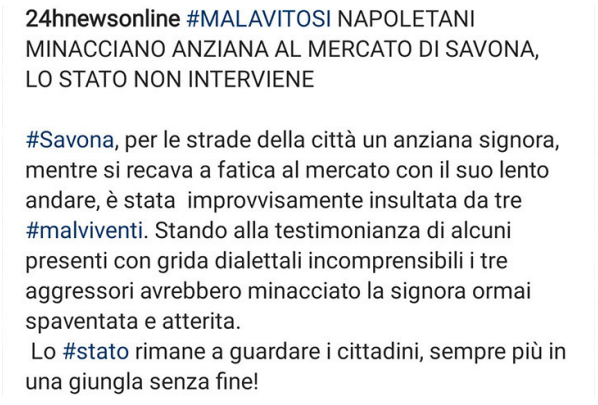
The annotation scheme (inspired by Bourgeade et al. (2023) and Cignarella et al. (2023)) includes different layers, i.e., **Stereotype**, **Stance** and **Forms of Discredit**. For **Stance**, the scheme used is well known in rumor detection literature (Aker et al., 2017), and includes the following four labels:

- **S** for **Support**: The comment supports the veracity of the story.
  e.g. *"Ormai è quasi quotidianitá, segno della grande mancanza di rispetto della maggior parte degli italiani. Maleducati e irrispettosi"*
  (transl. *"It's almost everyday life now, a sign of the great lack of respect of the majority of Italians. Rude and disrespectful"*)

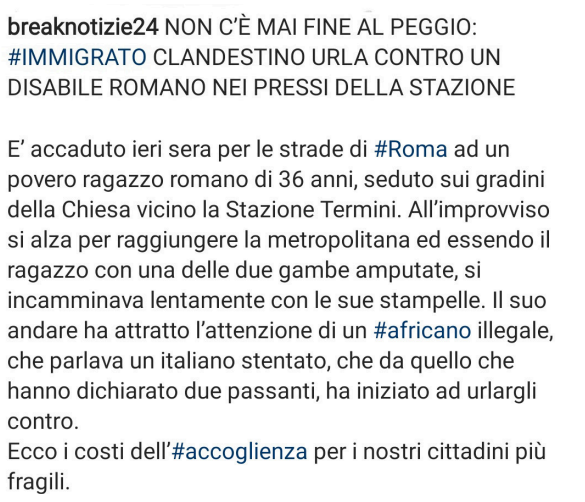
- **D** for **Deny**: The comment denies the veracity of the story.
  e.g. *"Fake news"*

- **Q** for **Query**: The comment questions the veracity of the story, requesting more information before making a judgement.
  e.g. *"Bisognerebbe analizzare i fatti accaduti piu approfonditamente, le notizie riportate sui social non sempre sono veritiere o riportate in maniera neutrale Per poterla descrivere dovrei conoscere il fatto accaduto approfonditamente da fonti certe"*
  (transl. *"We should analyze the events that happened in more depth, the news reported on social media is not always truthful or reported in a neutral manner. To be able to describe it I would have to know the event in detail from reliable sources"*)

- **C** for **Comment**: The comment does not directly relate to the veracity of the story.
  e.g. *"Lo stato italiano non ci aiuta molto riguardo alla malavita, soprattutto nei paesi del sud e gli anziani ne risentono ancor di più essendo loro più fragili..."*
  (transl. *"The Italian state doesn't help us much regarding crime, especially in southern countries and the elderly suffer even more as they are more fragile..."*)



(a) European hoax. (transl. *24hnewsonline #NEAPOLITAN CRIMINALS THREATEN AN ELDERLY WOMAN AT THE SAVONA MARKET, THE STATE DOES NOT INTERVENE. #Savona, on the streets of the city an elderly lady, while she was struggling to go to the market with her slow pace, was suddenly insulted by three #criminals. According to the testimony of some present, three attackers threatened the now frightened and terrified lady with incomprehensible dialect shouts. The #state remains watching citizens, increasingly in an endless jungle!*)



(b) African hoax. (transl. *breaknotizie24 THINGS CAN ONLY GET WORST: #ILLEGAL IMMIGRANTS SCREAMING AT A DISABLED ROMAN NEAR THE STATION. It happened last night on the streets of #Rome to a poor 36-year-old Roman boy, sitting on the steps of the Church near Termini Station. Suddenly he gets up to get to the subway and, being the boy with one of his two legs amputated, he walked slowly with his crutches. His walk attracted the attention of an illegal #African who spoke broken Italian, who, from what two passers-by said, started shouting at him. Here are the costs of #welcome for our most fragile citizens.*)

Figure 1: Fabricated racial hoaxes examples

For the **Stereotype** layer, the scheme distinguishes between the presence of explicit stereotypes, the presence of implicit stereotypes, and the absence of stereotypes of any kind. As identifying implicit expressions of stereotypes can be difficult, in this work we rely mainly on the criteria defined by Schmeisser-Nieto et al. (2022).

For this purpose, we adapted the scheme used in Schmeisser-Nieto et al. (2022), which individuates 13 linguistic indicators for the implicit and three for the explicit. In this article, stereotypes are classified as explicit when they refer to the nationality, origin and/or ethnic features of individuals or groups, including both cultural values and physical appearance. In addition, we characterized the stereotypes as explicit when occurring in copulative sentences, including cases of ellipsis of the copula, if used to confer offensive characteristics to individuals or groups. As far as implicit stereotypes are concerned, we adopted three of the linguistic markers used in Schmeisser-Nieto et al. (2022). Particularly: 1) the use of anaphoric expressions that refer to the target of the stereotype, which can also appear with omitted or vague expressions; 2) the human need to retrieve knowledge about events and facts from our shared knowledge of the world to understand the message 3) the use of figures of speech or irony in which the uttered message is different - and in some cases even opposed- to what the message actually conveys.

- **I** for **Implicit**.
  e.g. *"...Sono delle persone spregevoli che passano la vita facendo queste azioni, invece di andare a lavorare o rendersi uliti alla società"*
  (transl. *"...They are despicable people who spend their lives doing these actions, instead of going to work or serving society"*)
- **E** for **Explicit**.
  e.g. *"ORRIBILE E INCREDIBILE IGNORANTE, POCO RISPETTOSO"*
  (transl. *"HORRIBLE AND INCREDIBLE IGNORANT, NOT RESPECTFUL"*)
- **NO** for **No Stereotype**.
  e.g. *"...Se avesse compiuto il fatto il soggetto in questione ha sbagliato"*
  (transl. *"...If he had carried out the act, the person in question was wrong"*)

Finally, if a stereotype is present, it is annotated into one of six possible **Forms of Discredit** as described in Bourgeade et al. (2023) and inspired by the Stereotype Content Model introduced by Fiske (1998).

- **B** for Attack to the **Benevolence**.
  e.g. *"è ingiusto che una persona anziana o giovane che sia, debba essere derubata. Ladro"*
  (transl. *"It is unfair that an old or young person should be robbed. Thief"*)

- **AC** for **Affective Competence**.
  e.g. *"...Purtroppo per la poca moralità dell'immigrato non si può intervenirre ma spero che si faccia solo un esame di coscienza per aver insultato una persona fragile..."*
  (transl. *"...Unfortunately due to the lack of morality of the immigrant it is not possible to intervene but I hope that we just examine our conscience for having insulted a fragile person..."*)

- **C** for **Competence**.
  e.g. *"Il gesto compiuto è stato vergognoso. Senza cervello e arrogante"*
  (transl. *"The action taken was shameful. Brainless and arrogant"*)

- **DU** for **Dominance Up**.
  e.g. *"Cerchiamo sempre di aiutare qualsiasi persona, ma al momento del bisogno veniamo solo bullizati..."*
  (transl. *"We always try to help anyone, but when we need it we are only bullied..."*)

- **DD** for **Dominance Down**.
  e.g. *"...Un malvivente frustrato"*
  (transl. *"...A frustrated criminal"*)

- **P** for **Physical**.
  e.g. *"Orribile, aberrante."*
  (transl. *"Horrible, aberrant."*)

3.3. Annotation Process and Inter-Annotator Agreement

The annotation process involved three expert annotators, among which two female and one male. Each message was annotated for the categories and subcategories by two of the annotators, while the third intervened in the adjudication process to resolve disagreement and obtain gold labels for all the annotation layers, except for **Discredit**: for this subcategory, due to its very high subjectivity (as can be seen in Table 1) and also sparsity (typical of a multi-class category), we could not achieve a good agreement and thus preferred taking a more perspectivist approach, and thus kept both labels for each instance. We are planning a future extension of the corpus that will allow us a more reliable analysis of this category also. Figure 2 presents the distribution of annotated labels for each layer post-adjudication.

Table 1 presents the inter-annotator agreement pre-adjudication for each of the annotation layers. For the **Stereotype** category, we present the "strong" and "weak" agreements, respectively with and without considering the Implicit/Explicit distinction. For the **Discredit** subcategory, we also propose to collapse the 6 different classes into a reduced set of 4 (which group two pairs of often co-occurring forms of discredit), as well as a reduced set of 2 based on the *Agency* and *Warmth* concepts introduced by Fiske (1998).

As can be observed, the main **Stereotype** layer has a strong inter-annotator agreement, whereas **Stance** and **Discredit** appeared to be more subjective, and less balanced overall (as can be seen from Figure 2).

		Cohen's κ	IAA%
Stereotype	**Strong**	0.7963	90.32%
	Weak	0.8277	92.50%
Stance		0.5677	85.09%
Discredit	**6-way**	0.3422	51.16%
	4-way	0.3209	51.55%
	2-way	0.7882	56.59%

Table 1: Cohen's Kappa and percentage inter-annotator agreement for: the **Stereotype** dimension, with (**Strong**) and without (**Weak**) Implicit/Explicit distinction; the **Stance** dimension; the Forms of **Discredit**, in the original **6-way** (B,AC,C,DU,DD,P), collapsed **4-way** (B+DU,AC,C+P,DD), or **2-way** (*Agency*=C+P, *Warmth*=B+DU+DD).

4. Lexical Analysis

In Figure 2, we present the distribution of labels across each category, compared to the gold standard labels. These gold labels have been derived from the annotations of a third annotator, who resolved disagreements between the initial two annotators. It is important to note that the gold labels apply exclusively to the categories of Stereotype and Stance. For the category of Discredit, the situation is different. The Cumulative Discredit chart does not reflect a gold label standard but rather shows a cumulative and per-annotator distribution. This illustrates not only the overall frequency of discredit as identified collectively but also provides insight into the individual annotator's perspective on each form of discredit.

Table 2 provides a Lexical Analysis for the **Stereotype** layer, organized into three distinct categories: Explicit Lexicon, Implicit Lexicon, and No Stereotype Lexicon. Important keywords associated to

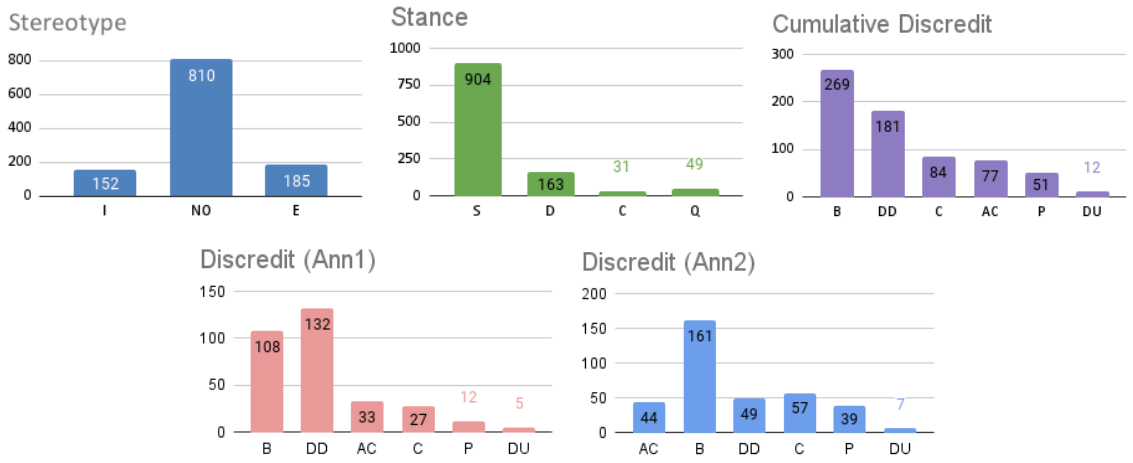


Figure 2: Distribution of labels for the different annotated dimensions. Forms of **Discredit** were not adjudicated, and as such are presented in cumulated and per-annotator forms (with the "None" class, corresponding to No Stereotype, excluded for clarity).

these classes are listed, alongside their corresponding TFIDF scores, which reflect their relative importance within the subsets of the corpus. In the Explicit Lexicon (a), words such as 'delinquent', 'criminal', and 'Neapolitan' feature prominently, with 'delinquent' having the highest TFIDF score of 14.94, indicating a strong association with explicit stereotypes. The Implicit Lexicon (b) contains words like 'uncivil' and 'educate', with lower TFIDF scores, suggesting a more subtle association with stereotyping. Lastly, the No Stereotype Lexicon (c) includes words like 'uncivil' (repeated with a higher TFIDF score here) and 'shame', which are significant yet not directly related to stereotyping, based on the context of the analysis. In Table 4 and Table 2, following the methodology outlined in Table 2, we have extended the lexical analysis to encompass the labels for 'Stance' and 'Form of Discredit'. This analysis maintains the use of TFIDF to quantify the significance of each lexicon within the respective categories. By applying this analytical approach, we aim to identify the most salient terms that are associated with each label, thereby providing a linguistic footprint of how different concepts are discussed within the dataset.

5. Linguistic Observations

Straying from quantitative analysis, we will now focus on a short selection of comments extracted from the corpus, which present linguistic phenomena well documented in literature and capable of conveying implicit messages. These are found especially in political propaganda and advertising language, but they are also well rooted in everyday language. In the following paragraphs, we will offer an analysis of the most noteworthy comments, which show phenomena such as presuppositions, implicatures and figurative language. For space reasons, we present only the translations of the messages.

1. *"I believe that this kind of news are widespread, especially in some areas of Italy with high crime rates. These news are really sad, but these events are very common. For sure, I wouldn't describe that person as they have been called in the comments, but no matter how this person has grown up, they committed a very serious action that deserves to be punished."*

It is interesting to observe how the quantifier "some" in the prepositional phrase "in some parts of Italy" generates two different implicatures. The first one is a classic case of scalar implicature, which seems to imply here that the author of the message is referring exclusively to some regions, and not to each of them. Scalar implicatures occur with expressions that signal a value within a scale and, when used, they usually imply the negation of the higher value of the scale (Bianchi, 2003). The second implicit meaning is more ambiguous and it concerns Grice's maxim of relation. The student seems to base his thought process on the stereotype according to which the alleged region of origin of the attackers (Campania) has a high level of crime, which would explain a piece of news like this. In doing so, however, the utterer ignores the fact that the event happened in Savona (Liguria), and not in Campania. We can thus see how the internalization of a stereotype can sometimes be misleading, even for the person who expressed it.

2. *"Like in the previous piece of news, we can see how the targeted victims are always frag-*

Explicit Lexicon	TFIDF
delinquente	14.94
criminale	6.86
napoletano	6.14
schifoso	3.48
vergogna	3.09

Implicit Lexicon	TFIDF
incivile	2.87
prossimo	2.41
educare	2.39
cercare	1.93
problema	1.91

No Stereotype Lexicon	TFIDF
incivile	11.60
vergogna	10.92
sapere	9.17
etnia	7.06
inaccettabile	6.89

Table 2: Lexical Analysis for **Stereotype**

DU Lexicon	TFIDF
persona	0.345
trovare	0.287
episodio	0.254
accadere	0.220
accadere società	0.220

DD Lexicon	TFIDF
malvivente	12.833
malavitoso	4.962
notizia	4.597
dovere	4.533
anziano	4.199

B Lexicon	TFIDF
delinquente	14.452
criminale	5.683
notizia	4.519
dovere	3.407
persona	2.989

C Lexicon	TFIDF
ignorante	2.364
ingiurre	0.912
anziano	0.879
anziano ignorante	0.879
volere	0.879

AC Lexicon	TFIDF
fidare	1.616
bisognare	1.576
persona	1.471
ragazzo	1.336
educare	1.234

P Lexicon	TFIDF
schifoso	1.970
schifo	1.954
schifo schifoso	1.665
animale	0.971
schifo animale	0.674

Table 3: Lexical Analysis for Form of **Discredit**

ile and weak people, in this case a guy in a wheelchair. Disgusting."

The author of the comment refers to a young disabled man, who experienced verbal aggression in the city of Rome, as a "guy in a wheelchair". This piece of information is not reported in the fake news, as the hoax article never states that the victim was in a wheelchair. The author of the message adds this false information without realizing it, operating on the false stereotype by which the prototype of "disabled person" is one who moves in a wheelchair. In this comment, the author uses – consciously or not – a synecdoche that conveys the message that moving in a wheelchair, while only being one of the many forms of disability, is enough to denote the whole category of disabled people.

3. *"I can't find the words to express the anger I feel towards these frequent episodes, even though the State decides to welcome those who are in pitiful conditions and especially to give them a job and better life conditions than the ones in their countries, they pay us back in this way…Obviously I am not painting everyone with the same brush, but the immigrants that pay respect and gratitude towards those who try to help them are fewer and fewer. I wouldn't even define them as human beings, but if I had to, I would say they're ungrateful people."*

In this comment, the author of the message used a fairly complex syntactic strategy to express a racial stereotype. First of all, they introduced a new, semantically vague referent with the referential expression: "those who find themselves in pitiful conditions". In doing so, he activated a presupposition and placed the referential expression in the position of the direct object of the complement clause, so that it was more difficult for a potential reader to argue its validity. In Italian, this syntactic position is usually occupied by old information, already known to those who participate in the speech situation, and being considered less salient from a cognitive point of view, it tends to go more unnoticed. Furthermore, the following anaphora related to the referent also occupies a similar role of direct object – usually, the referents in these positions have semantic roles that are not agentive. It is no coincidence that the anaphora covers this position when the author talks about the advantages that these people receive from the State. When the referent is later taken up anaphorically, the speaker shifts it into the syntactic role of the subject, which often coincides with the semantic role of agent, so as to be able to better indicate immigrants as those responsible for negative behavior.

Comment Lexicon	TFIDF
leggere	1.19
interessante	0.98
notizia descrivere	0.90
tema	0.72
importante	0.68

Support Lexicon	TFIDF
vergognoso	27.09
malvivente	20.78
schifo	19.28
orribile	19.15
ignorante	18.86

Query Lexicon	TFIDF
vero	1.67
condannare	1.35
urlare	1.35
cattivo	1.33
accadere	1.26

Deny Lexicon	TFIDF
aggressore	5.21
specificare	3.92
immigrato	3.83
odio	3.65
accadere	3.61

Table 4: Lexical Analysis for **Stance**

6. Discussion and Conclusion

The paper introduces a novel Italian corpus collected in the context of psychological experiments involving teenage students in schools. In this corpus, Stereotype, Stance, and Forms of Discredit were annotated. First of all, this corpus gave us the opportunity to study a text genre not often addressed in the literature about the detection of stereotypes and related phenomena, considering that the research community works mostly on social media platforms, which are not as frequently used by teenagers, at least in Italy. Secondly, we applied an annotation schema that takes into account a set of categories focused around the manifestations of stereotypes from the psychological literature, and we validated them by showing that they are lexically distinguishable in the analyzed comments. In future work, the annotation scheme applied to this corpus will be used in the annotation of a larger set of data and comparisons with other text genres will be developed. This will enable to expand upon the limits of this study and to collect more evidence about the validity of the categories that are applied in the annotation. We will also be able to exploit the unique characteristics of this data, to assist in the training of more robust stereotypes detection models.

Acknowledgments

The work of E. Chierchiello is funded by the International project *STERHEOTYPES - Studying European Racial Hoaxes and sterEOTYPES*, funded by Compagnia di San Paolo and VolksWagen Stiftung under the 'Challenges for Europe' Call for Projects (CUP: B99C20000640007).
The work of T. Bourgeade is funded by the project StereotypHate, funded by the Compagnia di San Paolo for the call 'Progetti di Ateneo - Compagnia di San Paolo 2019/2021 - Mission 1.1 - Finanziamento ex-post'.
The work of C. Bosco is partially funded by both the cited projects.

Limitations

The dataset used in this study was collected during 2022 and 2023 in a group of Italian schools. They are the outcome of an experiment conducted with small groups of students, whose attitudes can greatly vary over time. Therefore, the findings drawn from this dataset may not reflect the previous or future landscapes.

The dataset focuses specifically on Italian, limiting its generalizability to other languages and cultures. The sentiment about other people and the stereotypes triggered by the news created by psychologists for the experiment could be not representative of other set of teenagers.

The reduced amount of data is something that will be addressed in the future, but it is currently a limit of this preliminary work that mostly aims at providing a methodology to be tested in the future on larger datasets.

The limitations or biases arising from the dataset creation process, including data collection and annotation, should be considered in terms of the specific involvement of the annotators and the potential power dynamics that may have influenced the creation of the dataset.

Ethical reflections

As specified in the original publications pertaining to the source dataset (Corbelli et al., 2023; D'Errico et al., 2023), the student participants who produced the comments for these research projects were overseen by school staff, and appropriate informed consent forms were filled and signed by their legal guardians as necessary. No participation were refused or withdrawn, and an appropriate debriefing session was conducted after the last phase of the study. The Helsinki ethical principles and AIP (Italian Psychology Association) ethical code were followed, and the study was approved by the ethics committee of the University of Bari (reference code: ET-22-01).

The study presented in the paper can raise ethical considerations that should be carefully taken into account when collecting, analyzing and disseminating the data and results.

It is important to consider the possible misuse or unintended consequences of NLP tools. Care should be taken to avoid using systems that unintentionally and disproportionately target particular perspectives or promote misinformation on the raised issues. We can address this aspect by con-

sidering annotations even in disaggregated form, but a thorough analysis of the ethical implications of the tools developed should be conducted. Our work highlights the need to consider and incorporate the subjectivity of annotators in NLP applications and encourages thinking about the different perspectives encoded in annotated datasets to minimize the amplification of biases.

To ensure responsible and ethical use, we intend to implement mechanisms to track the use of the dataset. By recording who accesses and uses the dataset, we aim to promote a better understanding of its impact, encourage collaboration and potentially address concerns that may arise from its use. The dataset will be made available for research purposes only.

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39, Varna, Bulgaria. INCOMA Ltd.

Montserrat Nofre Mariona Taulè Enrique Amigò Berta Chulvi Paolo Rosso Alejandro Ariza-Casabona, Wolfgang S. Schmeisser-Nieto. 2022. Overview of detests at iberlef 2022: Detection and classification of racial stereotypes in spanish. *Procesamiento del Lenguaje Natural*.

Gordon Allport. 1954. *The Nature of Prejudice*. Routledge.

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*, pages 54–63.

Valentina Bianchi. 2003. *Pragmatica del linguaggio*. Laterza.

Cristina Bosco, Felice dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the evalita 2018 hate speech detection task. In *Proceedings of EVALITA '18, Evaluation of NLP and Speech Tools for Italian*, Turin, Italy.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *IberEval@SEPLN*. CEUR-WS.org.

Yi-Ling Chung, Aida Mostafazadeh Davani, Debora Nozza, Paul Rottger, and Zeerak Talat. 2023. Introduction to the proceedings of the 7th workshop on online abuse and harms (woah). In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*. ACL.

Alessandra Teresa Cignarella, Simona Frenda, Tom Bourgeade, Cristina Bosco, Francesca D'Errico, et al. 2023. Linking stance and stereotypes about migrants in italian fake news. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, volume 3596, pages 1–8. Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini, Nicole Novielli.

Giuseppe Corbelli, Paolo Giovanni Cicirelli, Francesca D'Errico, and Marinella Paciello. 2023. Preventing prejudice emerging from misleading news among adolescents: The role of implicit activation and regulatory self-efficacy in dealing with online misinformation. *Social Sciences*, 12(9).

Francesca D'Errico, Paolo Giovanni Cicirelli, Giuseppe Corbelli, and Marinella Paciello. 2023. Addressing racial misinformation at school: A psycho-social intervention aimed at reducing ethnic moral disengagement in adolescents. *Social Psychology of Education*.

Filippo Domaneschi and Carlo Penco. 2016. *Come non detto. Usi e abusi dei sottointesi*. Laterza.

F. D'Errico and C. Papapicco. 2022. 'immigrants, hell on board'. stereotypes and prejudice emerging from racial hoaxes through a psycholinguistic analysis. *Journal of Language and Discrimination*, (6):1–16.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing*

and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, pages 1–9. CEUR-WS.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@SEPLN*. CEUR-WS.org.

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

K.C. Fraser, S. Kiritchenko, and I. Nejadgholi. 2022. Computational modeling of stereotype content in text. *Frontiers in Artificial Intelligence*, 5:1–21.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

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

Mirko Lai, Fabio Celli, Alan Ramponi, Sara Tonelli, Cristina Bosco, and Viviana Patti. 2023. Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*.

Víctor Bargiela-Xavier Bonet Mariona Taulé, Montserrat Nofre. 2023. Newscom-tox: a corpus of comments on news articles annotated for toxicity in spanish. *Language Resources and Evaluation*.

Anne Reboul. 2011. A relevance-theoretic account of the evolution of implicit communication. *Studies in Pragmatics*, 13(1):1–19.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*, pages 1–9. CEUR-WS.

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

J. Vaes, M. Latrofa, C. Suitner, and L. Arcuri. 2017. They are all armed and dangerous! biased language use in crime news with ingroup and outgroup perpetrators. *Journal of media psychology: Theories, Methods, and Applications*, 31(31):12—23.

Teun A. van Dijk. 2016. Racism in the press. In Nancy Bonvillain, editor, *The Routledge Handbook of linguistic Anthropology*, chapter 25, pages 384–392. Routledge, New York.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çagri Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *CoRR*, abs/2006.07235.