Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09791-3

ORIGINAL PAPER

StereoHoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes

Wolfgang S. Schmeisser-Nieto · Alessandra Teresa Cignarella · Tom Bourgeade · Simona Frenda · Alejandro Ariza-Casabona · Mario Laurent, et al. ***[full author details at the end of the article]***

Accepted: 29 October 2024

Abstract

Stereotypes have been studied extensively in the fields of social psychology and, especially with the recent advances in technology, in computational linguistics. Stereotypes have also gained even more attention nowadays because of a notable rise in their dissemination due to demographic changes and world events. This paper focuses on ethnic stereotypes related to immigration and presents the StereoHoax corpus, a multilingual dataset of 17,814 tweets in French, Italian, and Spanish. The corpus includes conversational threads reporting on and responding to racial hoaxes about immigrants, which we define as false claims of unlawful actions attributed to specific ethnic groups. This work describes the data collection process and the fine-grained annotation scheme we used, which is based mainly on the Stereotype Content Model adapted to the study applied to immigrants of Bosco et al. (2023). Quantitative and qualitative analyses show the distribution and correlation of annotated categories across languages, revealing, for instance, intercultural differences in the expression of stereotypes through forms of discredit. To validate our data, we performed four machine learning experiments using pre-trained BERT-like models in order to lay a foundation for automatic stereotype detection research. Leveraging the StereoHoax corpus, we gained crucial insights into the importance of context, especially in relation to the detection of implicit stereotypes. Overall, we believe that the StereoHoax corpus will prove to be a valuable resource for the automatic detection of stereotypes regarding immigrants and the study of the linguistic and psychological patterns associated with their dissemination.

Keywords Racial Hoax · Immigration · Social psychology · Natural language processing · Stereotype detection · Corpus analysis

Warning: This paper includes examples that may contain instances of vulgarity, degrading terms and hate speech, which may be offensive to some readers.

Published online: 19 December 2024

1 Introduction

Categorization is one of the most fundamental strategies that people adopt to make sense of the world and other people. However, categorization is also a mechanism that can lead to the creation of stereotypes and promote prejudice toward other groups (Allport et al., 1954). In recent decades, stereotypes have been among the most important topics in social psychology, and renewed interest in ethnic stereotypes has been motivated, especially by the changing demographic characteristics and by the world events affecting contemporary societies. This includes a growing population of older people, wars and ethnic conflicts in several countries around the world, the subsequent intensification of migration flows, and the rise of extreme right-wing ideologies.

Language is undoubtedly the predominant means by which stereotypes are communicated in interpersonal discourse, and several studies on social categorization and language confirm that a thorough knowledge of stereotypes can be acquired through an analysis of linguistic data (Nicolas et al., 2021; Maass et al., 2006). The application of forms of automated analysis and computational linguistics approaches to different text genres, languages, and communication strategies are increasingly present in these studies: Sánchez-Junquera et al. (2021a), Ariza-Casabona et al. (2022) and Bosco et al. (2023).

In this paper, we describe a multilingual corpus in which ethnic stereotypes relating to immigrants have been annotated using a fine-grained scheme based on a well-known psychological framework. This corpus, hereafter StereoHoax,[1] is the result of a joint effort by the STERHEOTYPES project research group,[2] which involves researchers from Italy, France and Spain, all countries in which the phenomenon of ethnic stereotyping is indissolubly and historically linked to the arrival of immigrants from certain regions, mainly from Africa and Latin America.

The presentation of a preliminary release of this corpus was published in Bourgeade et al. (2023), which this paper extends upon in several key aspects: by discussing in greater depth the reference literature that comes from different disciplines, such as linguistics, computational linguistics and psychology; by reporting the results of novel statistical analyses and deep learning experiments; and by providing a more detailed description of the StereoHoax corpus itself.

Firstly, the originality of this corpus lies in the strategy applied for collecting data and the sources from which these data were drawn. Considering that stereotypes often occur in conversations in which misinformation is spread, we used debunking sites to collect a set of fake news texts based on false claims that a crime or unlawful action had been committed by a member of a specific ethnic group, or descriptions of threatening situations or characteristics related to that group, namely *racial hoaxes* focused specifically on immigrants. Using the items in this set as seeds, we collected conversations about these hoaxes from Twitter. By collecting reactions to these hoaxes from social media users, we gained access

[1] The StereoHoax corpus is available for research purposes upon request.

[2] https://www.irit.fr/sterheotypes/.

to the conversational context of the hoaxes. This methodology, though partially conditioned by the differences in the data arising from the cultural context of the three languages (Italian, French, and Spanish), allowed us to collect three comparable corpora, in which it is possible to observe the behavior of social media users regarding stereotypes related to immigration. Our specific focus on comments on Twitter responding to racial hoaxes was motivated by the very nature of the phenomena we wanted to study, that is, by the fact that stereotypes are very frequently and rapidly spread on social media such as Twitter.

Secondly, in order to go deeper into the features of these stereotypes in different cultural contexts, we created an annotation scheme inspired by the Stereotype Content Model (SCM) psychological framework (Fiske et al., 2007) and subsequent elaborations on the SCM (Abele et al., 2016; Koch et al., 2016; Fiske, 2018), and applied it to the StereoHoax corpus. Examples of the few datasets that can be used in computational linguistics experiments and in which stereotypes are annotated include Ariza-Casabona et al. (2022) and Sanguinetti et al. (2020), but these datasets come with annotation schemes that do not highlight the characteristics of stereotypes at a fine-grained level. Moreover, they do not account for the conversational dimension involved in the spread of stereotypes, ignoring the relevance that context may have in the detection of stereotypes. Our scheme identifies the presence of stereotypes, describes the associated forms of discredit, shows whether the stereotypes are implicit or explicit in the messages and whether their interpretation depends on the associated conversational context (composed of the racial hoax and reactions to it on Twitter).

Finally, we validated the annotation scheme by performing deep learning experiments which highlighted the importance of contextual information for the detection of stereotypes in conversations. In particular, we performed experiments using pre-trained BERT-like models, such as `CamemBERT`, `GilBERTo` and `BETO` for French, Italian and Spanish, respectively.

The StereoHoax corpus was created to improve the detection and classification of racial stereotypes related to immigration, for the first time taking into account the conversational context in which a message occurs and should be interpreted. It is devised to be a resource for training and testing machine learning systems with the novelty of making available key information regarding the context of the messages, thereby facilitating their classification. Composed of texts in different though comparable languages, the StereoHoax corpus is a starting point for the investigation of racial stereotypes in a multicultural and intercultural setting, and for identifying the psychological and linguistic patterns related to them. The results of the experiments performed confirm the usefulness of this dataset and of taking into account the context of the conversation in the annotation.

This paper is divided into seven sections. In Sect. 2, we present the theoretical grounds from the fields of social psychology and computational linguistics that underpin our work and discuss the definitions of the notions upon which this work is based, i.e., *stereotype*, *discredit*, and *hoax*. In Sect. 3 and Sect. 4, we introduce the StereoHoax corpus and the annotation scheme, respectively. Section 5 presents the statistical analyses performed on our data. In Sect. 6, we describe the experimental settings applied to our data, as well as an error analysis of the experiments. Finally, in Sect. 7, we discuss conclusions and future work.

2 Related work

2.1 Racial hoaxes

A **racial hoax** (RH) is a particular form of information disorder (Wardle, 2018) and can be defined as a communicative act featuring distorted and misleading information in the form of a threat to individuals' or societies' health and safety, in which the protagonist is a person, or a group of people described in terms of their ethnicity, nationality, or religion (Cerase & Santoro, 2018).

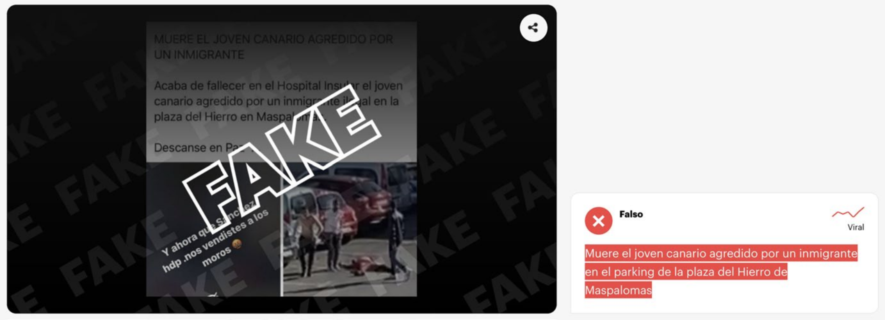
In a broad sense, racial hoaxes are examples of fake news based on false or biased claims that certain unlawful actions have been committed by a member of a specific ethnic group.

Racial hoaxes are expressed through linguistic distortions aimed at creating a perception of threat. They are transmitted primarily through biased words such as 'illegal immigrants' or 'COVID positive' or by means of a description of immigrants' actions, such as committing rape, spreading diseases, and obtaining social advantages (for concrete examples of racial hoaxes, see Sect. 3). In particular, a racial hoax frequently contains linguistic forms that are typical of stereotypes and prejudices (D'Errico et al., 2022), which can be a powerful way to propagate distorted information by reinforcing anti-immigrant attitudes (Wright et al., 2021). In order to define racial hoaxes or misleading news, it is crucial to take into consideration both the use of stereotypical and biased language and the untruthfulness of the news being reported (D'Errico et al., 2022).

Racial hoaxes can be seen as an additional form of news bias that can contribute to the spread of ethnic prejudice and discrimination (Esses et al., 2013), as already evidenced in crime news stories in which the use of biased language promotes associations between crime and immigrants (Vaes et al., 2017). A recent content analysis study conducted on a corpus of Italian racial hoaxes described their psycholinguistic characteristics (D'Errico et al., 2022), which were found to be based mainly on stereotypical language and simplistic content, thereby favoring heuristic and automatic processing of the news.

In general, immigrants are defined in RHs as perpetrators of negative actions, with a marked tendency for the "journalist" to adopt an aggravating and discrediting stance. Immigrants are rarely described as victims of negative actions and negative actions against them are often described in statistical terms, thereby fostering emotional detachment from them (Arcuri, 2015). An example of a fact-checked headline[3] is shown in Fig. 1.

[3] https://www.newtral.es/bulo-pelea-inmigrante-maspalomas-muere-joven-canario/20210121/.



MUERE EL JOVEN CANARIO AGREDIDO POR UN INMIGRANTE

Acaba de fallecer en el Hospital Insular el joven canario agredido por un inmigrante ilegal en la plaza del Hierro en Maspalomas.

Descanse en Paz

Y ahora que Sánchez hdp .nos vendistes a los moros

Falso

Viral

Muere el joven canario agredido por un inmigrante en el parking de la plaza del Hierro de Maspalomas

Fig. 1 Image from the fact-checking site 'Newtral,' showing a racial hoax circulated on social media. The headline translates to: 'Young Canary Islander attacked by an immigrant in the parking lot of Plaza del Hierro in Maspalomas dies'

2.2 Stereotypes and forms of discredit

According to Allport et al. (1954), a **stereotype** is the product of a categorization process consisting of attributing certain features to a social group, such as immigrants, who are described systematically, for instance, as *criminals* or *illegals*. These attributed labels are generalized to cover a whole set of people who share certain characteristics of identity, regardless of individual variations (Allport, 1935). Stereotyping is a rapid, automatic process that saves cognitive resources by creating a kind of shortcut by employing simplified structures, thereby reducing the amount of information to be considered. This process generates cognitive beliefs that are generally related to the content of the stereotype, e.g., *women cannot drive*, but also includes evaluative, motivational, and emotional elements that express a subjective stance toward a social group, which may become the object of prejudice and discrimination (Allport, 1935). Stereotypes can also be reinforced through specific linguistic choices.

From a psycholinguistic point of view, stereotypes can be expressed both implicitly and explicitly. For instance, they are implicit and subtle when the speaker exaggerates cultural differences or refuses to express positive feelings towards the target group (Pettigrew & Meertens, 1995). Implicitness can be conveyed by more complex forms such as metaphors (Collins & Clément, 2012) and negation bias (Beukeboom et al., 2010). More recently, Schmeisser-Nieto et al. (2022) drew up criteria for identifying implicit stereotypes by means of several linguistic strategies including anaphoric reference, rhetorical questions, irony or sarcasm and humor and jokes (Poggi et al., 2011a). In contrast, the explicitness of the stereotype can be recognized when the speaker directly attributes a negative evaluation of immigrants by means of derogatory adjectives or insults (Collins & Clément, 2012; D'Errico & Paciello, 2018; D'Errico and Poggi, 2014).

A closely related theoretical notion is that of dehumanization, which highlights how language can be used to express extreme forms of out-group denigration (Bandura, 2002) through linguistic forms that negate targeted human characteristics, such as the ability to reason or feel emotions (Haslam, 2006), thereby lowering humans to the level of animals or even to the level of viruses or diseases through the language of biological dehumanization (Utych, 2018).

In this sense, in addition to focusing on specific ways of expressing stereotypes, psychosocial literature has also highlighted the content of stereotypes, which can be associated with different types of *prejudice* (Fiske et al., 2007; Bye, 2020). In an evaluative process based on underlying stereotypes and prejudice, a negative conclusion involves discrediting the image of the other. Several forms of discredit have been identified in the literature (Poggi et al., 2011b) based on the Stereotype Content Model (SCM) proposed by Fiske et al. (2007).

Based on SCM and subsequent adaptations (Fiske et al., 2007; Fiske, 2018), the two macrocategories of discredit content are called *competence* and *warmth*. However, regarding the terminology of the macrocategories, we adopted the names proposed by Abele et al. (2016), *agency* and *communion*, agreed later by Fiske (2018) to be parallel to competence and warmth.

These terms can be applied to several types of stereotypes, such as gender, sexual orientation and age, but in our examples, we have focused on racial stereotypes, specifically those related to immigrants. Previously, Lee and Fiske (2006) conducted a study on the perception of immigrants in the United States using the SCM dimensions of competence and warmth. The study found that the generic image of immigrants is perceived as low on both dimensions. However, when specific immigrant groups determined by their nationality and socioeconomic status were considered, perceptions tended to be ambivalent: that is, while some groups were perceived as highly competent but lacking warmth, others were seen as high in warmth but low in competence.

Following Abele et al. (2016), in the case of communion, individuals may be associated with a lack of qualities that are related to the establishment and maintenance of social relationships, such as morality, benevolence, fairness, and kindness. At the same time, for agency, the stereotypical content refers to a lack of qualities that are related to individual achievements, such as being capable, intelligent and skilful (Ybarra & Stephan, 1994; Abele et al., 2016; Fiske, 2018).

In our project, the theoretical model taken as reference is, in the first instance, the seminal one of Fiske (Fiske et al., 2007; Fiske, 2018), integrated later with the model of discredit (Poggi & D'Errico, 2010) and the ABC (*A*gency/Socioeconomic Success, Conservative-Progressive *B*eliefs, and *C*ommunion) model (Koch et al., 2016). The latter two models explicitly stress the importance of including the subcategory of 'dominance' in the social evaluation of groups. In addition to the theoretical foundations, the application of a study focused on the detection of ethnic stereotypes on social media (Bosco et al., 2023) proved to be highly valuable, as it provided a corpus of field data for the attribution of stereotypes.

The integration of theoretical and applied studies allows us to refer to these two macrocategories of communion and agency, and to recognize their subcategories. Communion includes *affective competence*, *benevolence* and *dominance up*, while agency includes lack of *competence*, *physical* and *dominance down* (Poggi & D'Errico, 2010; Abele et al., 2016; Koch et al., 2016; Bosco et al., 2023), as shown in Table 1.

Table 1 Adapted stereotype content model with two macrocategories and its subcategories

Communion	Agency
Affective competence	Competence
Benevolence	Physical
Dominance up	Dominance down

Particularly, in the case of communion, as shown in the work aimed at detecting stereotypes in social media regarding immigrants (Bosco et al., 2023), contents related to benevolence present actions or characteristics pointing to a lack of trustworthiness, honesty, ethics and legality. Therefore, the lack of benevolence is associated to descriptions of them as thieves, rapists, murderers, criminals or exploiters. Within the macrocategory of communion, authors also found that immigrants were described in terms of a negative exercise of dominance over groups or societies (Cheng et al., 2013; Poggi et al., 2011b) and, in this case, the immigrant is frequently described as dangerous, overbearing, generally aggressive and arrogant, even symbolically (Castelfranchi, 2023), with statements such as ”*they want to impose their values and customs*” or “*they come here and want/aspired to luxury*” (Bosco et al., 2023; D’Errico et al., 2022). This description meets the definition of dominance up in the model of discredit (Poggi & D’Errico, 2010; Castelfranchi, 2023), which is also generally associated with social or group harm.

The other subcategory associated with communion is affective competence, in which the emotionality of the individuals follows the description given by Bosco et al. (2023), that is, to evaluate immigrants only negatively and as “emotionally inadequate”. For instance, immigrants may display a strong emotional reaction when they are angry or, in the opposite case, they can be sad and suffering. In the first case, they show an emotion with high arousal (Russell, 1998), as in the case of exaggerated anger, expressed in their difficulty in channeling their emotions, while in the second case they express emotions with low arousal, as in the case of excessive sadness.

On the other hand, within the macrocategory of agency, which focuses more on the individual characteristics of the target, there is the subcategory of competence, in which the individual is accused of lacking intellect and being ignorant, stupid, unprepared and incomprehensible (Fiske, 2018; Nicolas et al., 2021). For example, this can be seen in the case of immigrants when the messages allude to their lack of intelligence or education. In contrast, the physical subcategory, which is not present in the work of Fiske (2007; 2018), highlights unpleasant physical characteristics of the target group. Within this framework, immigrants are described as dirty, sick, ugly and physically disgusting, as well as carriers of diseases (Marshall & Shapiro, 2018; Bosco et al., 2023; D’Errico et al., 2022).

Finally, from the results of Bosco et al. (2023), it emerged that a lack of agency may correspond to a lack of generic power, or a lack of assertiveness in the case of (Nicolas et al., 2021), whose subjects are described as dependent and inactive. This lack of power, labeled as dominance down (Bosco et al., 2023; Poggi & D’Errico, 2010; Castelfranchi, 2023), includes negative labels, such as being a parasite, a do-nothing, a complaining slacker or someone who passively takes advantage of the system (D'Errico et al., 2022).

2.3 Stereotypes in computational linguistics

Although discourses containing stereotypes and prejudices have always existed, with the rise of social media, the spread of stereotypes has been amplified and reinforced. In order to tackle this issue, in recent years, the computational linguistics community has worked to develop resources and systems to recognize stereotypes in different types of discourse, including political debates (Sánchez-Junquera et al., 2021a), online newspaper comments (Ariza-Casabona et al., 2022) and Facebook posts (Bosco et al., 2023). Since the source can influence linguistic differences between texts, we highlight the importance of making datasets available from a variety of social media platforms.

Recent works have focused on the type of narrative in which stereotypes are present, of which racial hoaxes are an example. For instance, Card et al. (2016) clustered profiles of *latent personas* based on similarities and evaluated this model in news articles about immigrants. On a similar note, Fokkens et al. (2018) proposed the extraction of *microportraits*, a collection of descriptions in a text used to refer to an entity, in this case, Muslims. Beukeboom and Burgers (2019), as well as Sap et al. (2020), used the concept of frames to study how biases in language transmit stereotypes, which are expressed implicitly. Other techniques used for detecting stereotypes in texts are word representations (Bolukbasi et al., 2016), a text masking technique and BETO, a BERT-based model for Spanish (Sánchez-Junquera et al., 2021b) and techniques for natural language inference (Dev et al., 2020).

With regard to Fiske's SCM (Fiske et al., 2007), Fraser et al. (2022a) proposed a computational model for categorizing stereotypes in two macrocategories: *competence* (*agency*) and *warmth* (*communion*).

Several other papers in this area of research consider this theoretical model. For example, Vargas et al. (2023) develop an approach called Social Stereotype Analysis, which consists of analyzing stereotypical beliefs by contrasting them with counter-stereotypes. Ungless et al. (2022) show that the SCM model is particularly able to capture biases in contextualized word embeddings, offering a mitigation procedure based on them. The SCM has also been used to detect various forms of stereotypes in large-scale language models, e.g., in the context of sensory impairment (Herold et al., 2022) and intersectional features (Cao et al., 2022) in pre-trained language models, which also show how language models treat physical and mental illness (Mina et al., 2024). The stereotypes associated with age in social media texts and psychological literature, on the other hand, are the focus of Fraser et al. (2022b, 2024).

Bosco et al. (2023) and Bourgeade et al. (2023) extended the categories based on this psychological model into six subcategories: *affective competence, benevolence, dominance up, competence, physical* and *dominance down*. Another work that presented a taxonomy of stereotypes about immigration is Sánchez-Junquera et al. (2021a), in which immigrants are perceived as *victims*, as *resources* or as a *threat*. Based on this scheme, Ariza-Casabona et al. (2022) these categories are, in turn, subcategorized into nine topics following the content with which the stereotype is associated. Since stereotypes are either latent or manifest in language, the presence of biases in language models is not unfamiliar, and numerous efforts have already been made to mitigate them, some of them based on well-grounded stereotype theories as Koch's ABC (Cao et al., 2022).

Other works on stereotypes about immigrants were presented in evaluation shared tasks, such as DETESTS at IberLEF 2022 (Ariza-Casabona et al., 2022) and HaSpeeDe 2 at EVALITA 2020 (Sanguinetti et al., 2020). Regarding stereotype detection, several shared tasks targeting women have been presented, such as AMI held in both EVALITA 2018 (Fersini et al., 2018) and EVALITA 2020 (Fersini et al., 2020) campaigns, and the EXIST task, which was held at the IberLEF 2021 (Rodríguez-Sánchez et al., 2021) and IberLEF 2022 (Rodríguez-Sánchez et al., 2022) campaigns. Chiril et al. (2021) presented a dataset annotated with gender stereotypes tested on deep learning models.

In conclusion, the present work makes several significant contributions to the study of ethnic stereotypes related to immigration, starting from the basis of state-of-the-art work on social psychology, linguistics, and natural language processing. First, it introduces the StereoHoax corpus, a multilingual dataset consisting of 17,814 tweets in French, Italian, and Spanish, which is specifically focused on racial hoaxes about immigrants. Second, the paper details the data collection process and presents a fine-grained annotation scheme based primarily on the SCM by Fiske et al. (2007). Third, through both quantitative and qualitative analyses, the authors reveal the distribution and correlation of annotated stereotype categories across different languages, uncovering intercultural differences in stereotype expression. Fourth, the paper validates the dataset by conducting four machine learning experiments using pre-trained BERT-like models, paving the way for future research in automatic stereotype detection. Finally, the authors emphasize the importance of conversational context in detecting implicit stereotypes, making the StereoHoax corpus a valuable resource for studying linguistic and psychological patterns in the dissemination of stereotypes.

3 The StereoHoax corpus

The creation of the multilingual Stereotypes and Hoaxes (StereoHoax) corpus is a joint effort by a multidisciplinary research group made up of social psychologists and computational linguists. In this section, we provide an overview of the characteristics of StereoHoax (Sect. 3.1) and present the three main steps involved in the creation of this corpus, whose interest lies particularly in conversation threads that emerge from RHs (Sect. 3.2). The examples from this section onwards have been taken from StereoHoax and translated into English to guarantee anonymity.

3.1 Overview of StereoHoax

The StereoHoax corpus comprises 17,814 tweets, of which 9342 are in French (referred to as StereoHoax-FR); 3123 are in Italian (referred to as StereoHoax-IT); and 5349 are in Spanish (referred to as StereoHoax-ES). Figure 2 shows the percentages of tweets by language within the corpus. The starting point for their selection is a collection of RHs published and refuted on fact-checking websites in France, Italy and Spain.

3.2 The creation of StereoHoax

There were three main steps in the creation of the StereoHoax corpus, (i) the selection of racial hoaxes (Sect. 3.2.1), (ii) the definition of strategies for the selection of the tweets from the Twitter API V2, (iii) and the pre-processing of the tweets to form a unified multilingual dataset that takes into consideration ethical issues and possible biases (Sect. 3.2.2).

3.2.1 Collecting racial hoaxes

We started the process by manually collecting a set of 239 racial hoaxes (70 in French, 97 in Italian and 72 in Spanish) that were related to the issue of immigration from French, Italian, and Spanish fact-checking websites or newspaper articles verifying or refuting claims made in social media. For French, we used *AFP Factuel*

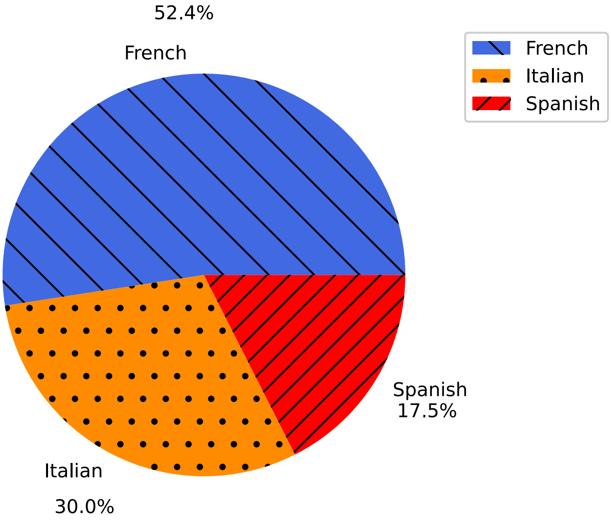


Fig. 2 Distribution of language data in the StereoHoax corpus

and *Les Décodeurs* in *Le Monde*; for Italian, *Bufale.net* and *Butac*; and for Spanish, *Maldita.es* and *Newtral*.[4] The debunked RHs dated between 2019 and 2022. Our focus is on the hoaxes more widely discussed.

Consequently, we identified the words in the texts denoting the target group to ensure that the hoaxes were related to immigrants and we then classified the racial hoaxes into five categories in accordance with the main topic they addressed. This classification was inspired by the taxonomy proposed by Sánchez-Junquera et al. (2021a) and Ariza-Casabona et al. (2022). Initially, Sánchez-Junqueras identified three macrocategories, in which immigrants are perceived as *Victims*, an *Economic Resource* or a *Threat*. Since one of the characteristics of RHs is to highlight a negative feature of a social group, we applied to our collection the subcategories related to *Threat* in addition to an extra category *Others*.

These subcategories are described below and associated with examples (translated into English, with underlined target groups and words expressing subcategory main concepts).

- *Security:* events related to safety, such as thefts, public disorders, fights, sexual assaults, murders, and terrorist attacks. Example (1) links Moroccans, i.e., the target group, to a criminal action such as starting a fire.

  (1) *Moroccans who arrived illegally on the coast of Ceuta have assaulted the Juan Morejón School and have started a fire.*
- *Public health:* health issues that may potentially affect the population, mainly infectious diseases such as COVID-19, HIV, Ebola and Malaria. In Example (2), the target group is associated with COVID-19, however, the racial hoax also overlaps with the category of security, when it mentions a riot or hostages.

  (2) *Riot in a reception center in Rome, migrants with COVID take hostages: "We want to get out."*
- *Migration control:* events related to migratory flows, arrivals, disembarkations, border control and the regulation of immigration, as seen in Example (3), in which immigrants outnumber the previous inhabitants of a city.

  (3) *In Fuerteventura and Lanzarote, the number of Moors is already greater than that of the autochthonous population.*
- *Benefits:* situations in which immigrants are perceived as receiving more help, social assistance and welfare benefits than non-immigrants. The implication of Example (4) is that those 1100 million euros are at the expense of nonmigrants. Here, there is a certain overlap with the category of migration control when the target is referred to as *illegal*.

  (4) *Paying for health care for illegal immigrants costs 1100 million euros.*
- *Culture and religion:* cultural and religious differences are perceived as threatening the traditions of nonimmigrants. This can be seen in Example (5), in which Islam is perceived as a foreign and imposed religion.

[4] https://www.lemonde.fr/les-decodeurs/; https://www.bufale.net/; https://www.butac.it/; https://maldita.es/; https://www.newtral.es/.

Table 2 Percentages of types of RHs in the three language subsets

Language	Benefit (%)	Security (%)	Migration control (%)	Public health (%)	Religion (%)	Others (%)	Total
Italian	4.12	58.76	15.46	20.62	0.00	1.03	97
Spanish	29.16	25.00	16.66	12.50	13.88	2.77	72
French	50.00	25.00	19.44	0.00	5.56	0.00	70

(5) *The Balearic Islands want to eliminate the Catholic religion in schools and they want Islam to be taught in some centers.*

- *Others:* in Example (6), there is a racial hoax targeting immigrants that does not appear in the previous categories, in which immigrants are described as being instrumentalized for the benefit of politicians.

(6) *The left supports immigrants to gain votes for the next elections.*

One expert per language (two psychologists for Italian and Spanish and a linguist for French) assigned a RH category to each racial hoax. The RH categories correspond to a subcategory of immigrants seen as a threat (Sánchez-Junquera et al., 2021a) and a set of keywords related to the topic and associated with the target group, such as benefits, tradition, assault or borders. Although rare, a racial hoax may contain more than one category. In those cases, the experts labeled the racial hoax taking into account the main message according to their interpretation of the main communicative intention of the authors. For instance, in Example (4), the main focus is the money invested in immigrants' health care, although the irregular legal situation of the migrants, expressed by the use of the adjective illegal, is also present. Table 2 shows the distribution of racial hoaxes by topics and by language. For a more detailed analysis of the distribution of these RH categories with respect to the categories of discredit, see Sect. 5.

3.2.2 From racial hoaxes to reactions

The next step was to devise strategies to find and extract tweets related to the collected racial hoaxes. In our collection of racial hoaxes, we saved the URLs from which they were extracted. The URLs contained certain features that we subsequently applied to retrieve the tweets, for which we used *Twitter API v2* for Academia.[5] The fact-checking websites we used had different formats and features, therefore we needed to apply different strategies to retrieve the relevant tweets. These strategies were:

[5] At the time of collection the name of the social media was Twitter, while is currently X, and allowed the use of the following API for data collection: https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries/v2.

(a) Tweets embedded in the fact-checking website. In this case, the fact-checker reports the RH in a tweet. We applied this strategy to fact-checkers in French and Spanish.
(b) Quoted text within the fact-checking website, which would generally contain the core raw text of the RH. This feature was helpful when extracting tweets in Spanish.
(c) Searching for and matching the headline of the RH in the Twitter API. This strategy proved useful for finding tweets containing RHs in Italian.
(d) Fact-checking websites with Twitter accounts. With Twint, a Python Twitter-scraping tool, we searched for tweets containing the word ‘migrant’ in those Twitter accounts and then manually checked whether those tweets were in response to a tweet propagating a racial hoax. This strategy was applied for retrieving tweets in French.

After retrieving the tweets, we expanded the search to extract the whole conversation thread using the conversation ID provided by Twitter. The initial retrieval resulted in 9342 instances for the French dataset, 13,200 for Italian, and 165,621 for Spanish. To reduce noise within the Italian and the Spanish datasets, we applied two data cleaning strategies. By doing so, we ensured a higher probability of tweets containing stereotypes, with or without the target group being explicitly expressed. The strategies are as follows:

- for Italian, we removed all the duplicates that had no replies, leaving in the dataset only those tweets, even duplicates and retweets, that generated conversations
- for Spanish, we removed retweets and duplicates, and we kept only those tweets containing keywords referring to the target group found in racial hoaxes, such as *immigrant*, *illegal*, *Muslim* and *black*, as well as their parent and child tweets (regardless of whether they contained keywords).

After data cleaning and data annotation of the Italian and Spanish datasets, the size of our initial dataset decreased to 3123 and 5349 tweets, respectively. Finally, text pre-processing of the selected tweets was performed. Usernames were removed to preserve anonymity. Emojis were translated into their official CLDR textual form (e.g., `:smile:`), using the `emoji`[6] Python library, and URLs were replaced with a `[URL] title-of-the-website [\URL]` format. Hashtags were left unchanged. The data collection process is shown in Fig. 3.

4 Annotation of the StereoHoax corpus

In this section, we introduce the hierarchical annotation scheme designed and applied to manually annotate the StereoHoax corpus. We present the annotation scheme in two-level categories, as well as the theoretical background associated

[6] https://pypi.org/project/emoji/.

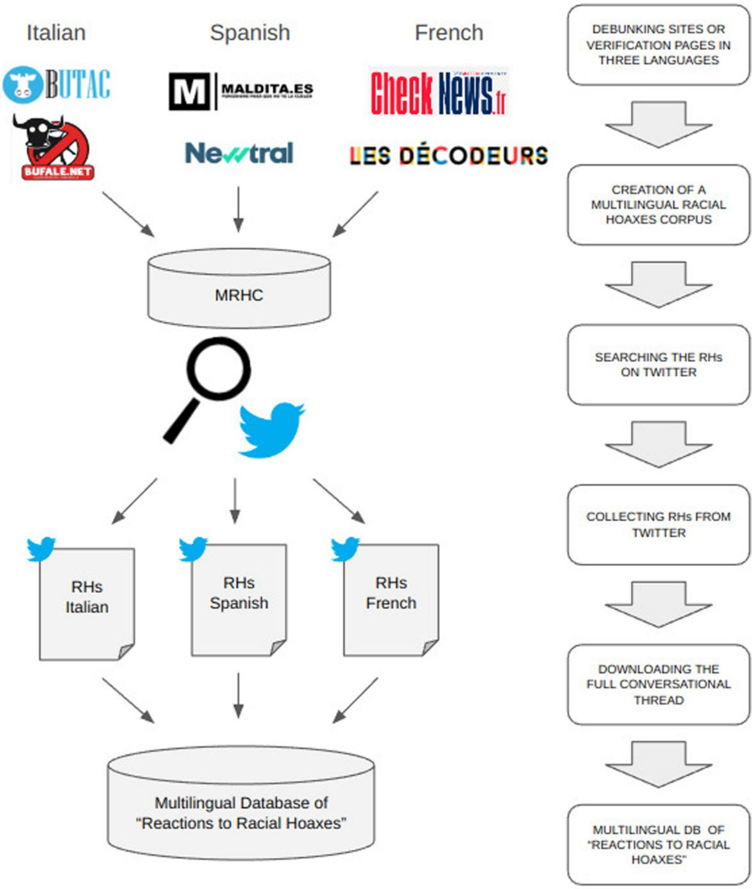

Italian
Spanish
French
BUTAC
BUFALE.NET
MALDITA.ES
Newtral
CheckNews.fr
LES DÉCODEURS
MRHC
RHs Italian
RHs Spanish
RHs French
Multilingual Database of "Reactions to Racial Hoaxes"
DEBUNKING SITES OR VERIFICATION PAGES IN THREE LANGUAGES
CREATION OF A MULTILINGUAL RACIAL HOAXES CORPUS
SEARCHING THE RHs ON TWITTER
COLLECTING RHs FROM TWITTER
DOWNLOADING THE FULL CONVERSATIONAL THREAD
MULTILINGUAL DB OF "REACTIONS TO RACIAL HOAXES"

Fig. 3 Data collection process. This figure has been extracted from Bourgeade et al. (2023)

with it (Sect. 4.1). Consequently, we describe the annotation process for StereoHoax, including the inter-annotator agreement (IAA) tests applied (Sect. 4.2).

4.1 Annotation scheme

A high level of subjectivity is one of the biggest challenges in accomplishing the task of annotating stereotypes and classifying them consistently. Therefore, the creation of clear and non-ambiguous guidelines, which limit the proliferation of interpretations by the annotators, is fundamental for the accomplishment of this complex task. Therefore, we formulated guidelines defining categories by integrating the classical approach to stereotype content (Fiske et al., 2002, 2007; Fiske, 2018) with the dimension of dominance or power (Poggi et al., 2011b; Koch et al., 2016) in its active (dominance up) and passive (dominance down) forms, since the latter study showed a strong association between these categories and immigrants in evaluations

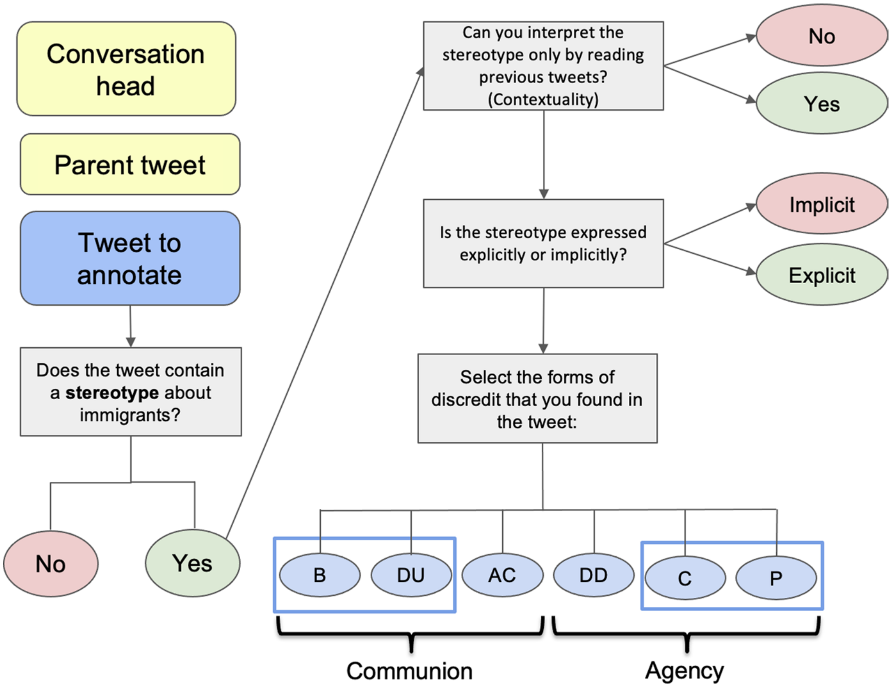

Conversation head
Parent tweet
Tweet to annotate
Does the tweet contain a **stereotype** about immigrants?
No
Yes
Can you interpret the stereotype only by reading previous tweets? (Contextuality)
No
Yes
Is the stereotype expressed explicitly or implicitly?
Implicit
Explicit
Select the forms of discredit that you found in the tweet:
B
DU
AC
DD
C
P
Communion
Agency

Fig. 4 Diagram of the annotation tool display. The initials B, DU, AC, DD, C and P correspond to the six forms of discredit: Benevolence, Dominance Up, Affective Competence, Dominance Down, Competence and Physical

of this social group on social media (D'Errico et al., 2022; Bosco et al., 2023). We also considered different ways of expressing stereotypes, i.e., whether they were expressed explicitly or implicitly and whether the context from previous tweets was needed to understand the stereotype contained in the message (see Sect. 2). The annotation scheme consists of four layers organized in two hierarchical levels. Figure 4 illustrates the diagram of the hierarchical levels of the annotation layers and their answers followed by the annotators.

Level 1: presence of stereotypes The first level refers to the presence of stereotypes in the tweet. This category has a binary value with *yes*/*no* indicating the presence or absence of a stereotype, respectively. In the guidelines, a stereotype was defined as:

- *Stereotype*: a negative or positive[7] feature is associated with a target group, in this case, immigrants. At a cognitive level, the feature is applied homogeneously to the entire group, which is often described in terms of its members' place of origin, ethnicity, or religion. An association is created between the members of the group and characteristics related to crime, welfare benefits, taxes, diseases,

[7] In StereoHoax, the presence of positive stereotypes is unusual because the annotated tweets are derived from RHs.

employment, cultural differences and their quality as human beings. This association, i.e., the stereotypes, can be expressed explicitly or implicitly. In Example (7), the elements presented in the definition of stereotype can be observed. The target group is manifested with the pronoun *they*, whose antecedent is the word *immigrants* (elided in the main tweet, but displayed in the example in square brackets), recovered from the previous tweet (the parent tweet or the conversation head), and the associated idea is that they will claim more territories (expressed explicitly) and that they are patient (expressed implicitly through the metaphor of the drop of water).

(7) *They [immigrants] will come here to claim more "territories", little by little, without haste. They do not care about time. They are like a drop of water that does not stop falling, that is, they fall permanently and eventually pierce a 1-meter-thick steel plate.*

Level 2: forms of Discredit, Contextuality and Implicitness The second level of annotation is considered when the stereotype is annotated as being present. This level of annotation comprises three categories: discredit, contextuality and implicitness, described in the guidelines as:

- *Discredit*:

  The forms of discredit are based on the so-called 'big two' macrocategories, communion and agency, which emerged from several psychosocial models (Fiske et al., 2007; Fiske, 1998; Abele et al., 2016), and later adapted following the works by Bosco et al. (2023); Poggi and D'Errico (2010); Koch et al. (2016). This is a multi-class label that includes six categories of discredit: benevolence, dominance up and affective competence, which correspond to the macrocategory of communion; and dominance down, competence and physical, which correspond to the macrocategory of agency. Since a tweet may contain more than one stereotype, it was possible to select more than one category. The definition of these six categories are the following:

  - *Benevolence (B)*: this category indicates that there is an attack on the benevolence and morality of the non-migrants. In the text of the tweet, the target portrayed as a criminal displays violent behavior or does not share a certain code of behavior or certain values related to morality. For instance, the immigrant is portrayed as a rapist, thief, terrorist, murderer or drug user). Example (8) refers to an immigrant as illegal, thereby attributing to them characteristics that contravene social norms, and Example (9) depicts them as criminals.

    (8) *And then, how do you integrate an illegal immigrant without documents?*

    (9) *First they take drugs, then they kill our young people so they are immediately released.*

  - *Dominance up (DU)*: this category is assigned when the target of the stereotype exercises a type of dominance and maintains it through arrogant or demanding behavior, for instance, when being overbearing, demanding a house or social services, or taking someone's job. In Example (10) and (11), the former suggests that immigrants will displace the majority religion in Italy with their own, and in the latter, that without strict laws, immigrants would 'invade' the country.

(10) *If you do nothing, Italy will be a Muslim country in ten years.*
(11) *AustraliaOnFire, with its very strict immigration laws, is immune from the barbarian invasion.*

- *Affective competence (AC)*: this category is assigned when there is an attribution of distorted or unregulated emotional behavior, both in the case of high arousal, as in Example (12), and in the case of low arousal, where immigrants are depicted as victims of discrimination or excessive suffering, receiving a paternalistic attitude, as seen in Example (13).

(12) *There is always an excuse: deranged, suddenly crazy.*
(13) *We must pity them, they are depressed and in difficulty, let's help them immediately.*

- *Dominance down (DD)*: this category is assigned when the target exhibits passive behavior and demands or receives something they do not deserve. The target is compared to a parasite or a slacker who always complains. The category is also used when migrants are dehumanized. Example (14) envisions a situation in which immigrants receive money without having earned it, whereas in Example (15) immigrants are dehumanized by being called parasites.

(14) *It is happening in Italy!! The card for migrants arrives: they can withdraw up to 37 euros a day!*
(15) *Yet it's the truth and then you're fed up with entertaining these parasites who have nothing to do with us.*

- *Competence (C)*: this category is assigned when the target is seem as cognitively inferior, unable to do skilled jobs, ignorant or stupid. In Example (16), it is stated that immigrants lack professional qualification. This example could also fit in the dominance down category, since it can be inferred that the immigrant has a job that they do not deserve since they do not have the necessary qualifications. On a similar note, Example (17) also portrays immigrants as less competent, but the inferiority of their competences is emphasized through ethnic differences, associating skin color with intelligence levels.

(16) *In a normal, democratic country, this type of person [an immigrant] would be in the unemployment queue for not having any type of professional qualification.*

(17) *Geniuses have no color, but it must be conceded that lots of individuals of a certain color are far from geniuses!*

- *Physical (P)*: this category is assigned when the target has specific unpleasant physical features and poses a dangerous threat to public health. The target is associated with infectious diseases or with being ugly or dirty. Example (18) expresses a preconceived idea that the author has on how Syrians should look. Therefore, this is an example of a stereotype based on physical appearance. From another perspective, Example (19) depicts immigrants as carriers of diseases.

(18) *Syrians? I see mostly black people.*

(19) *And of course, we don't take into account the effect of 'not treating a contagious disease, and the immigrant ends up becoming patient zero.*

Since the aim of this study is to ensure that all forms of discredit are represented, the six categories were grouped into four final categories, which allowed us to gather data in a more aggregated way for our computational goals. In particular, we decided to merge benevolence and dominance up into a single category since both categories evoke a harmful and potentially violent characterization of immigrants, from both a real and symbolic point of view. Both the competence and physical subcategories represent two sides of the same coin, with the former describing the cognitive characteristics of an individual and the latter describing physical ones.

- *Contextuality*: this is a binary category in which the annotator is given the options to respond *yes/no* as to whether contextuality is needed to interpret the stereotypes. It is described in the guidelines as follows:

  - *Yes*: to understand the meaning of the stereotype expressed in the tweet, you need to look at the context, i.e., previous messages in the thread or the RH that triggered it, or external elements such as websites and images. In Example (20), as well as in Examples (7) and (16), the previous tweets are needed to recover the antecedent from the anaphoric references *these ones*, *they* and *this type of person*, respectively. The antecedent of *these ones* in Example (20), *the migrants of the Aquarius*, is manifested in the parent tweet.

(20) <u>*These ones*</u> *don't look exhausted and sick!!! I would even say in great shape to get the allowances...!!!*
Parent tweet: *All this human suffering of [the migrants of the Aquarius] when they arrive in Valencia!*

  - *No*: all the information that you need to understand the stereotype is available in the text of the tweet, as seen in Example (21).

Table 3 Inter-annotator agreements in terms of Kappas (♣ Fleiss'; ◇ Cohen's)

Layer	Italian♣	Spanish♣	French◇
Stereotype	0.48	0.76	0.73
Contextuality	0.55	0.49	0.63
Implicitness	0.54	0.15	0.64

(21) *Immigrants rape a 16-year-old Italian girl, then kill her. African parasites are an insult to civil society.*

- *Implicitness*: this is a binary category in which the annotator needs to decide whether the stereotype expressed in the message is implicit or explicit. The criteria proposed by Schmeisser-Nieto et al. (2022) are given to the annotators and a general instruction is provided in the guidelines:
  - *Implicit*: to access the stereotype expressed in the text, you must make an inference. In Example (22), the underlying stereotypes are expressed indirectly. Firstly, by emphasizing the nationality of a person, in this case, Senegalese, generalizing, in that way, the situation expressed in the text to the entire group. Secondly, through a rhetorical question, which implies that immigrants are trying to impose their own modes of behavior. Thirdly, by expressing an exhortation involving the immigrant, whose message is motivated by the belief that they should not be "here".

    (22) *And who does this Senegalese think he is to say what we should do in Spain? Let him go to his country*

  - *Explicit*: the stereotype is expressed explicitly in the text of the message. When there are two or more stereotypes in the tweet, at least one of which is explicit, the tweet must be annotated as explicit. In Example (23), the author of the tweet uses descriptions of the target group that are clear and direct and the stereotypes that the author of the text is spreading must therefore be considered explicit.

    (23) *What is dangerous is the brainless immigrants, who misinterpret Islam and who kill innocent people, they are racist and dangerous!!!*

Table 4 Label distribution for the categories annotated in the StereoHoax corpus. The numbers in the last four columns (forms of discredit) do not add up to 100% because discredit could be annotated with more than one label per tweet. That is, tweets could be counted more than once

	Stereotype (%)	Contextuality (%)	Implicitness (%)	Forms of discredit			
				AC (%)	B + DU (%)	C + P (%)	DD (%)
Italian	15.11	43.22	70.34	4.24	81.78	26.06	16.10
Spanish	29.97	36.81	21.46	9.67	64.13	2.68	47.66
French	12.07	72.52	88.83	8.60	73.05	3.81	55.76

4.2 Annotation process

The annotation of the corpus was carried out simultaneously by the teams in France, Italy and Spain, who were in charge of the subsets in their respective languages (French, Italian and Spanish). The annotators were trained following the guidelines described in Sect. 4.2. The three teams annotated the tweets using Label Studio,[8] an open-source annotation tool. A common template was translated into the respective languages and was used for displaying tweets and questions to assign the annotation labels. The display for each instance of annotation included the conversation head, the parent tweet and, highlighted, the tweet to be annotated. The annotation procedure started with the first question regarding the presence of stereotypes: (1) Does the tweet contain a stereotype about immigration? If the annotator answered yes, then, the remaining questions were presented: (2) Can you interpret the stereotype only by reading previous tweets? (Contextuality); (3) Is the stereotype expressed explicitly or implicitly? (World Knowledge, Figures of Speech, Evaluation, Individual, Perpetrators, Exhortations), and (4) Select the forms of discredit you found in the tweet. In Fig. 4, there is a detailed diagram of the annotation hierarchy of questions as they appear in Label Studio.

Once the annotations were completed, we performed inter-annotator agreement tests, which were validated with different methods due to the different annotation procedures used for the different languages. Table 3 presents the kappa measures for each language and the stereotype, contextuality, and implicitness annotation labels. Due to the different processes of annotation adopted for each language, we used Cohen's kappa for French and Fleiss' kappa for Italian and Spanish.

The annotation of the French language texts was carried out in two stages. Firstly, two annotators trained for the task annotated small subsets of data and ran Cohen's Kappas coefficient tests to validate the annotation and readjust differences of criteria found in the disagreements. Secondly, two annotators were trained taking into account the new agreements reached by the team. Due to the sheer volume of data, each annotator annotated 50% of the data, with two cross-annotated sets of 300 and 144 tweets to ensure consistency. All the annotators of the French subset, one male, and three females, were French native speakers between 20 and 35 years of age,

[8] http://labelstud.io/.

ranging from undergraduate students to a Ph.D. in linguistics. In Table 3, we only present kappas for the final cross-annotated set.

For Italian, the whole dataset was annotated by two annotators, one male and one female, both native Italian speakers between 25 and 30 years of age, one of whom was a master's student in linguistics, while the other had a PhD in digital humanities. When the annotation was completed, three computational linguistics researchers, all native Italian speakers, females, with ages between 30 and 60 years, annotated the disagreements.

The Spanish team was made up of three annotators, two males and one female, between 23 and 40 years of age. One of them was a PhD student in computational linguistics, while the other two were undergraduates in linguistics who had been trained for the task. All of the annotators were native Spanish speakers. All three annotated the entire dataset, holding weekly meetings to solve questions regarding the guidelines. Disagreements were solved by majority vote, and for the cases in which all three annotators chose different options, the PhD student had the final say.

As we can see in Table 3, there are notable differences in the inter-annotator agreement scores for the different languages, even though the annotators were trained following the same guidelines. The main reason for this variation is the highly subjective nature of the analyzed phenomenon. The notable difference in the IAA score on contextuality and implicitness achieved by the Spanish annotators and the annotators from France and Italy is also due to the type of data which, as shown later in Table 4, tended to be more explicit in the case of Spanish.

5 Annotation analysis

In this section, we provide qualitative and quantitative analyses from a comparative perspective of the three-language subsets that make up the multilingual StereoHoax corpus, focusing in particular on stereotypes and forms of discredit in conversations, and the interaction of discredit with the other annotated dimensions.

5.1 Distribution of labels in StereoHoax

The newly released (compared to the version reported in Bourgeade et al. (2023)) StereoHoax is made up of 17,814 instances, of which 9342 tweets correspond to StereoHoax-FR, 3123 tweets to StereoHoax-IT and 5349 tweets to StereoHoax-ES. In Table 4, we report the distribution and percentage of each annotated category per language in line with the annotation scheme described in Sect. 4.1.

As can be seen, stereotypes are found more rarely in the Italian and French data than in the Spanish subset, which contains approximately 30% of the total number of stereotypes. Another commonality between the Italian and French subsets is the distribution of implicitness (70.34 and 88.83%), which is decidedly lower (21.46%) in the Spanish subset. In contrast, the Spanish subset contains a higher percentage of explicit stereotypes. Regarding contextuality, French is the language subset with

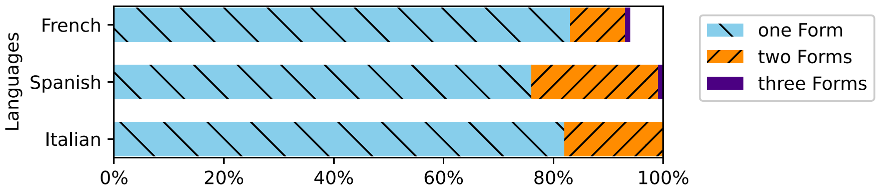


Fig. 5 Percentage of numbers of labels annotated regarding the forms of discredit (per tweet) in the three languages of the StereoHoax corpus

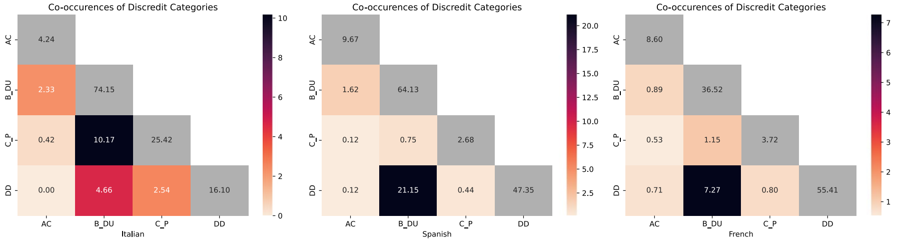


Fig. 6 Co-occurrence of forms of Discredit

the highest percentage of stereotypes that rely on context (72.52%), while Italian and Spanish both have percentages of approximately 40%.

It is interesting to observe some parallels to Lee and Fiske (2006). Although the American context in which their research was conducted differs in terms of the origins of immigration and the varying perceptions of ingroups, a similar pattern emerges for certain groups of immigrants, such as Arabs and undocumented immigrants. In Lee's study, Arabs, referred to as Muslim and North African immigrants in our dataset, are perceived as average in competence (agency) and low in warmth (communion). Similarly, undocumented immigrants are viewed as low on both dimensions. These findings align with our results, where the category of benevolence-dominance up (communion) has the highest percentages, followed by dominance down (agency). For instance, common stereotypes in our dataset include portrayals of immigrants as either terrorists or welfare beneficiaries, corresponding to the aforementioned categories.

Finally, the distribution of forms of discredit is similar in the French and Spanish subsets. In these two subsets, stereotypes are mainly concerned with the provision of social and economic aid by governments (dominance down), as well as criminality, illegality and fear of invasion (benevolence-dominance up). In Italian, the presence of the latter form of discredit is greater, followed by discredit regarding the competencies of immigrants and their physical attributes (competence-physical).

It is worth mentioning that more than one form of discredit can be present in a single tweet and, in some cases, more than one label has been annotated. Figure 5 shows the multi-label distribution of forms of discredit, i.e., the percentages of the number of different forms of discredit annotated by instance in each subcorpus. It can be observed that the percentages are consistent across languages: the majority of tweets containing a stereotype, a figure that exceeds 75% in each language, presents

Table 5 Detailed percentage distribution of RH categories with respect to forms of discredit

		Benefits (%)	Security (%)	Migration control (%)	Public health (%)	Religion (%)
Italian	AC	0.00	3.18	–	1.06	–
	B+DU	0.00	36.44	–	**37.71**	–
	C+P	0.00	9.96	–	15.47	–
	DD	0.21	8.47	–	7.42	–
Spanish	AC	6.49	1.19	1.75	0.00	0.25
	B+DU	**22.65**	11.79	10.36	0.00	19.34
	C+P	0.56	0.87	0.19	0.00	1.06
	DD	17.72	7.11	10.54	0.06	11.92
French	AC	6.12	1.68	0.80	–	–
	B+DU	18.09	14.10	4.34	–	–
	C+P	2.22	1.06	0.44	–	–
	DD	**52.04**	1.60	1.77	–	–

The numbers that are underlined and in bold represent the highest percentages for each language

only one form of discredit (blue bars), while between 9 and 25% of the total tweets contain two forms of discredit (orange bars). Three or more forms of discredit (purple bars) in the same tweet were found less than 1% of the times in the three languages. One notable case in the dataset, in which all four forms of discredit were found, is Example (24):

(24) *They have hatred*$^{B+DU}$*, they say, they are useless*^{DD} *who live in the vegetative*^{AC} *and in the animal instinct*$^{C+P}$*, pests*$^{B+DU}$*. We must treat them as such and not forget all those who brought this vermin*$^{B+DU/C+P}$.

Note that, in French, “pests” (*nuisibles*) and “vermin” (*vermine*) carry different meanings: the former is more “formal”, whereas the latter is much more negatively emotionally loaded.

5.2 Distribution of multiple forms of discredit

In order to verify how the labels representing the four forms of discredit are distributed across the tweets of the StereoHoax corpus, we measured the co-occurrence of such forms and present the results in Fig. 6. The three *heat maps* refer, from left to right, to the Italian, Spanish and French subcorpora, respectively.

According to the annotation scheme presented in Sect. 4.1, more than one form of discredit can be encountered within the same tweet. Therefore, in this paragraph, we highlight meaningful co-occurrences of forms of discredit in each language setting. As can be seen in Fig. 6, in Italian, the strongest co-occurrence is between the benevolence-dominance up and competence-physical categories, followed by the co-occurrence of benevolence-dominance up and dominance down. This second observation is also valid for the Spanish and French subcorpora, though with different intensities (see the different scales in the three heat maps). Example (25), which was extracted from the StereoHoax corpus, contains both the benevolence-dominance up and dominance down categories.

(25) *Of course they are not punishable they have nothing they are not prosecuted and they do whatever the fuck they want*$^{B+DU}$ *simple and then the police are in deep shit and they don't give a damn and they even mock us and the crazy thing is that we economically sustain them too*^{DD} *thanks #governmentofinjustice*

It is interesting to note that the data shows the frequent presence of two forms of discredit together, benevolence and dominance down, which according to the model proposed by Fiske (see Sect. 2.2) fall into the same macrocategory, i.e., communion. From a computational perspective, this can be considered to support an annotation scheme based on the macrocategories, which can be, in effect, more suitable when the fine-grained forms of discredit are sparse as in our dataset.

5.3 Forms of discredit versus RH topics

Another analysis that was carried out is related to the co-occurrence of the four different forms of discredit within each category of racial hoaxes described in Sect. 3.

Table 5 shows the percentage of tweets that contain the four forms of discredit: affective competence (AC), benevolence-dominance up (B-DU), competence-physical (C-P) and dominance down (DD) for each subcorpus. The first thing to note is that Spanish is the only language among the three considered in this study to display values for all categories of racial hoaxes, even though public health is almost non-existent in Spanish. In Italian, the values for the categories migration control and religion are absent, while in the French subset, the categories public health and religion are unrepresented. We attribute this disparity in distribution to the different cultural contexts in the three countries, which produce different racial hoaxes and, hence, different stereotypes. Below is a representative example extracted from the StereoHoax corpus in which the user mentions two recurring topics: 'immigration/conquering' and 'religion'.

(26) *We can't stop saying it and SEEING it: the immigration we have in front of us is not Vikings or those of the Rising Sun*🧐😤, *wherever they come from the south of the Mediterranean, it is with this conquering religion as a standard*🤮🤬💣

Secondly, from the column relative to the RH category of benefits, it can be observed that there is a higher co-occurrence with the forms of discredit benevolence-dominance up and dominance down in both Spanish and French. The Italian subset has very few representative items for this category. On the other hand, the two categories of racial hoaxes that are mostly represented by forms of discredit in Italian are security and public health, which co-occur mainly with the benevolence-dominance

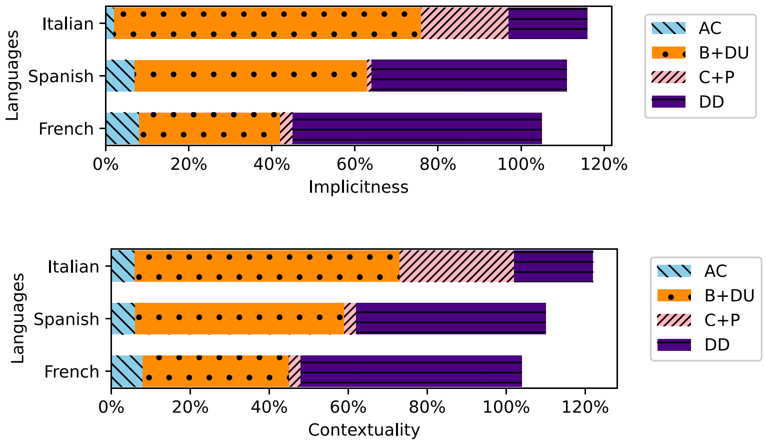


Fig. 7 Co-occurrence between implicitness/contextuality and forms of discredit. The stacked bar-charts do not add up to 100% due to the potential for a single tweet to bear multiple discredit labels (in which case it is counted the corresponding number of times)

Table 6 Prediction error distribution of RoBERTa-BNE models with hard and soft labels and GPT-4 according to the categories of implicitness and context of stereotyped instances

	Only conversational heads		Only reactions		Conversational heads and reactions	
	Cramer's V	χ^2/p-value	Cramer's V	χ^2/p-value	Cramer's V	χ^2/p-value
French	0.00	0.00/1.00	0.04	12.20/**0.02**	0.05	20.48/**0.00**
Italian	0.00	0.00/1.00	0.16	45.95/**0.00**	0.02	1.26/0.87
Spanish	0.06	5.67/0.23	0.14	211.11/**0.00**	0.14	254.19/**0.00**

The numbers in bold indicate significant correlations (p-value < 0.05) between the categories of implicitness and contextuality

up form of discredit. See, for instance, Example (27), in which the increasing spread of Coronavirus is blamed on migrants.

(27) *Infected immigrants, reception center becomes RED ZONE: "Health bomb with 1000 guests" [URL] We anticipated it yesterday: the boats brought the coronavirus back to Sicily. And now it expands.*

5.4 Forms of discredit versus implicitness and contextuality

In Fig. 7—above, we display the percentages of the co-occurrence of implicitness and the four different forms of discredit, while in Fig. 7—below, we show the percentage of the co-occurrence of contextuality and the four forms of discredit. In both images, the bar charts are grouped by language: Italian, Spanish and French and normalized by percentage.

As can be seen, contextuality predominantly co-occurs with the benevolence-dominance up form of discredit in all three languages. In Italian, the second most commonly co-occurring form of discredit with contextuality is competence-physical, while the most common for both French and Spanish is dominance down.

From these analyses, we draw the following conclusions: first, we observe a significant correlation between implicitness and contextuality in many configurations, with these aspects being heavily influenced by the specific categories of discredit involved in instances containing stereotypes. This observation validates our decision to merge certain discredit categories for analysis and experimental purposes since they exhibit similar trends in all three languages (namely, high degrees of contextuality and implicitness in benevolence and dominance up forms of discredit, with comparatively lower degrees in competence and physical forms). The results presented in Table 6 further highlight the necessity of examining the processing of contextuality in machine classification experiments, given its critical role in effectively dealing with implicitness in the detection of various forms of stereotypes in different segments of the datasets (such as conversation heads versus responses).

5.5 Conversational analysis

In this section, we provide an overview of the cross-relations between the different tiers of annotation in the multilingual StereoHoax corpus (implicitness, contextuality and forms of discredit).

We studied the associations in three different scenarios by considering tweets containing: (i) only the conversational heads, (ii) only the 'reactions' without the conversational heads and, (iii) the conversational heads and the reactions.

In Table 6, we display the values of Cramér's V, χ^2 and the p-value to confirm or disregard the association between the categories of implicitness and contextuality. The results show that this association is not significant (p-value > 0.05) in the first scenario (only conversational heads). Indeed, what can be observed by crossing the values of implicitness and contextuality is that, in general, stereotypes in conversational heads require the context to be explicit in very few cases.

Secondly, by looking at the second scenario in Table 6 (only reactions), it can be observed that implicit stereotypes expressed within the textual thread require the previous context to be inferred, i.e., significant at < 0.05. However, given the distribution of implicit and explicit stereotypes in the Spanish data, tweets containing stereotypes appear to be more explicit and do not seem to require as much context (for comparison, see the percentages in Table 4).

Finally, for Spanish and French, the trend of correlation between implicitness and contextuality is confirmed, while in Italian the majority of texts containing stereotypes are precisely the conversational heads themselves.

In this section, we have described and discussed the characteristics of the annotation of the multilingual StereoHoax corpus. In the following section, we will describe how we took advantage of these annotated data to perform machine learning experiments to validate the dataset and the annotation scheme.

6 Experiments and validation

In this section, we present some experiments useful to validate the three subsets of StereoHoax by language, together with results obtained and an error analysis.

6.1 Experimental setting

To validate our annotation scheme and establish baseline performances for stereotype detection models trained on our data, we made use of state-of-the-art NLP classification models. As this was the first set of experiments performed on this data, we chose to only consider the mono-lingual settings, leaving the exploration of multilingual aspects to future work.

Similarly, we chose to leave the classification of discredit categories to future work, as it is a heavily imbalanced multi-label prediction task, and preliminary experiments showed it to be very challenging when tackled with traditional methods. Therefore, we selected three existing appropriate pre-trained BERT-like models to fine-tune our data on:

- For French, we chose the `CamemBERT` (Martin et al., 2020) model, a `RoBERTa` architecture that was pre-trained on French subsets of a variety of large text corpora (OSCAR(Suárez et al., 2019), CCNet (Wenzek et al., 2020), Wikipedia).
- For Italian, we chose the `GilBERTo`[9] model, which was inspired by `CamemBERT` and based on `RoBERTa` on the Italian subset of the OSCAR corpus (Suárez et al., 2019).
- For Spanish, we selected the `BETO` (Cañete et al., 2020) model, based on BERT (Devlin et al., 2019), which was pre-trained on a mixture of Wikipedia and OPUS Project (Tiedemann, 2012) texts in Spanish, which the authors consider to be an updated version of the Spanish Billion Words Corpus (Cardellino, 2019).

To experiment with the contextual aspects of our data, we designed four different experimental settings, which varied in terms of the amount and nature of the contextual information given as input to the models: in *without (w/o) Context* setting, we only provided the main tweets' texts; in *with (w/) Parent* setting and *w/ Head* setting we additionally provided the text of the parent tweet or conversation head, respectively; and in *w/ Both* setting, we concatenated the previous two elements of context and provided them both at once.

To construct our datasets' train/test/validation splits, independently, for each language, we generated 20 stratified-grouped 5-folds (to get as close as possible to an 80–20% train-test split), with grouping by conversation-id to prevent breaking

[9] https://github.com/idb-ita/GilBERTo.

Table 7 Overview of datasets splits

		Train	Test	Validation
Italian	Size	1,841	1,185	97
	Size %	58.95	37.94	3.11
	Stereotype %	16.46	12.91	16.49
Spanish	Size	4,085	1,049	215
	Size %	76.37	19.61	4.02
	Stereotype %	29.52	31.84	29.30
French	Size	6,981	1,993	368
	Size %	74.73	21.33	3.94
	Stereotype %	12.02	12.29	11.96

Table 8 Evaluation results for our baseline experiments, in terms of F1-scores (both per-label, and macro-averaged), over the four contextual settings

Language-model	Class	Setting			
		w/o Context	w/Parent	w/Head	w/Both
IT-GilBERTo	Stereo	57.14	**65.08**	54.44	58.30
	Not-stereo	93.60	**95.85**	91.84	95.67
	Macro	75.37	**80.46**	73.14	76.98
ES-BETO	stereo	**75.52**	74.18	75.25	74.24
	Not-stereo	**89.61**	86.05	87.42	86.48
	Macro	**82.57**	80.12	81.33	80.36
FR-CamemBERT	stereo	50.20	45.67	49.02	**50.84**
	Not-stereo	92.82	**94.26**	92.52	92.38
	Macro	71.51	69.96	70.77	**71.61**

Best scores for each row are highlighted in bold

conversation boundaries across train-test splits, and stratification on the stereotype labels. Each of the 100 generated split solutions was then evaluated in terms of two factors:

1. the absolute difference in size of the test-set with 20% of the total language-set size, to ensure the splits were relatively equally sized (in proportion), per language;
2. the Wasserstein distance between the distributions of conversations sizes (in terms of the number of tweets) between the two splits, to ensure a relatively equal variety of larger and smaller conversations within the training and test sets.

These two factors were min-max normalized ($x, y \in [0;1]$) independently for each language (except for Italian, for which the split-size factor was weighted eight times more than the Wasserstein distance, after initial tests which yielded splits too different from 80 to 20%). The split with the smallest 2-norm in this 2-dimensional space was then selected as the "best" for the language. The validation set, due to its smaller size and for the sake of simplicity, was stratified as ~ 5% of the training set. Table 7 presents the effective sizes of these splits.

To ensure greater fairness in the cross-lingual comparison, we used the `Base` variant of each of the three models, as it was the largest (in number of parameters) common model-size available. Hyperparameters (learning-rate and effective batch-size, with gradient accumulation) were automatically fine-tuned per-language using Weights &Biases' Bayesian search method (Biewald, 2020), in the *w/o Context* setting only, to speed up the search.

6.2 Results

The evaluation results for our experiments can be found in Table 8, expressed as per-label and Macro F1-scores. We can first observe significant differences in performances between the three different languages. This was to be expected, as the language-specific subsets (and data splits) are not equally sized nor balanced concerning the stereotype labels. Fundamental differences are also to be expected in the phenomena effectively annotated due to differences in the cultural-linguistic contexts and subjectivity in the annotation criteria, among other potential factors. Additionally, we observed that each of the language-specific models behaved differently depending on the availability of contextual information: for example, for Spanish, the *w/o Context* setting seems to have yielded the best performance, whereas for Italian, providing the parent tweets led to significant improvements in the classification scores.

6.3 Error analysis

To qualitatively evaluate the mispredictions produced by our different models, we extracted all test-set predicted labels, together with their associated probabilities (obtained from the logits of the models), for each of the 12 trained models. Since each of the four variants for each language may have classified each instance correctly or incorrectly independently, we examined the predictions for all test-set instances. We observed different types of errors, depending on which contextual variants had misclassified given instances.

For the models applied to the *StereoHoax-IT corpus*, we notice that out of the 132 misclassified instances, 82 belonged to the following conversational head:

(28) *What was the request by the governors of the northeast to keep* ***Chinese children*** *under control for a period before letting them go back to school? Deeply #racist, right? In fact, now the schools are closing them. #scoundrels*

The above tweet is one of the few instances in the dataset that is not related to migrants that arrive on boats in the Mediterranean but rather manifests a stereotypical attitude towards another target in bold. Reviewing the textual content of the thread originated from the conversational head in Example (28), it emerges how, even though the head itself contains a racial hoax and a stereotype regarding a racially connotated target, the message is actually a criticism of politicians, especially left-wing parties, and of their management of the COVID-19 health crisis in 2020. Such texts clearly do not contain racial stereotypes, e.g., "*The virus arrived flying in first class with a white middle-aged Italian manager*", but the language model seems to be incapable of distinguishing between them.

In the *StereoHoax-ES corpus*, out of the 45 misclassified instances, 29 were predicted to contain no stereotype when the context was not taken into account. However, that number decreased dramatically when the models were trained with the conversation head, the parent tweet or both, with seven, eight and eight incorrect predictions, respectively. One of the patterns we observed in the misclassified tweets refers to the focus of the message. The topic of the tweet is not focused on immigrants, but the stereotypes appear when they are mentioned tangentially, normally as a cause or as a consequence of the issue under discussion. As we see in Example (29), the message is directed towards politicians. The author complains about their decisions, which benefit immigrants to the detriment of others. However, this misprediction is corrected when the models that handle the conversational head, the parent tweet (in this case, both correspond to the same tweet) and both contexts are used. In this case, the tweet refers to an agreement made between Spain and Morocco.

(29) *They are destroying our pensions and everything that our elders had achieved... Having to work so that immigrants live and we have to live poorly for all the governments we have had.*

Regarding the categories, most mispredictions when recognizing stereotypes are with the benevolence-dominance up subcategory, with eleven errors, and the dominance down subcategory, with six errors, of which four are cases of double labeling as benevolence-dominance up and dominance down. Given that these are the most prevalent categories in the Spanish subset, with 64.13% of benevolence-dominance up and 47.66% of dominance down, these results are not surprising. Although, category affective competence accounts for only 9.67% of the dataset, there were six misclassified instances, the equivalent of 21% of the mispredictions. Since most of the cases annotated as affective competence portray immigrants as victims, the dataset is likely biased regarding this representation.

For the *StereoHoax-FR corpus*, we observed that out of the 80 instances misclassified as having no stereotypes (out of 245 positive gold labels) by all variants, 39 belonged to the same conversation, which concerns a racial hoax about migrants supposedly "desecrating" a famous French basilica by their presence there. Examining the content of the relevant target tweets, we find that most stereotypes are of an abstract nature, often referring indirectly to the mere migrants' presence, as an attack on the authors' faith. Since we cannot find similar types of stereotypes in the French training set itself, we believe the models may have all struggled with this peculiar mode of expression. Similarly, since, as can be seen in Tables 4 and 5, the presence of competence-physical discredit is relatively low in the French subset, the four variants also had problems handling types of stereotypes based on general appearance and competences.

In Italian, of the 87 instances (138 for French; 69 for Spanish) which were correctly classified by at least one of the context-aware variants (*w/Parent/Head/Both*) but misclassified by the *w/o Context* variant, 21 (39; 53) were annotated as containing stereotypes, of which 18 (29; 24) were labelled by human annotators as requiring context to be interpreted, compared to three (10; 29) in which that was not the case. This shows that the need for contextual information was identified within the Italian and French subsets and that classification models do seem to benefit from having access to this information when it facilitates the interpretation of an instance.

The reasons for the limited benefits, or even adverse effects, specifically for Spanish, of incorporating contextual information into predictive models are unclear. However, the existing literature contains similar inconclusive findings regarding the integration of context into abusive language classification and related phenomena (Pavlopoulos et al., 2020; Cercas et al., 2021; Menini et al., 2021; Markov & Daelemans, 2022). One plausible hypothesis is that the addition of a vast number of context tokens might act as a "distractor" for Transformer architectures, leading to diminished performance. This aspect could be investigated in future work.

7 Conclusions and future work

In this paper, we have presented StereoHoax, the first multilingual and multilayer annotated corpus consisting of French, Italian and Spanish tweets, which are reactions to racial hoaxes related to immigrants. The annotation includes the identification of stereotypes and their classification into four forms of discredit, condensed from the original six categories. It also includes whether the stereotypes are expressed explicitly or implicitly, and whether their interpretation requires contextual information. Therefore, to annotate contextual information accurately, the conversational thread in which the tweet occurs is also included. This work extends on and enhances previous work (Bourgeade et al., 2023) by describing the first outcomes of a study of stereotypes that are spread through racial hoaxes, with the aim of creating NLP resources and tools for their automatic detection.

In order to address this challenging task, we started with a review of the psychological and computational literature regarding RHs and stereotypes. This helped us to build a collection of racial hoaxes in French, Italian and Spanish, classified according to the topic of the news topic they addressed.

We then presented a complete description of the multi-layered scheme for the annotation of racial stereotypes in social media data based on the Stereotype Content Model, already applied in a previous work (Bosco et al., 2023), in three different languages. We applied it, for the first time, to a newly created multilingual dataset of Twitter reactions to racial hoaxes, the StereoHoax corpus. Therefore, this work can be considered pioneering and the multi-layered annotation scheme might undergo a certain degree of adaptation when applied to datasets with very different characteristics. From a theoretical point of view it is necessary to underline that the annotation carried out is in line with the work of Bosco et al. (2023) in the sense that the categories detected include also the dominance subcategory as noted by Koch et al. (2016) and by Poggi and D'Errico (2010). Furthermore, unlike the classic work of Fiske (2007; 2018), essentially negatively oriented stereotypes emerged. Naturally, this can be due to the type of data, which strongly differs from the 'bottom-up' evaluation considered in psychosocial research, as opposed to the dataset extracted from social media, which can be strongly negatively characterized. This imbalance of positive/negative stereotypes in our dataset may represent a limitation in this present work. Therefore, in future studies, we will aim to build a dataset with a greater positive stereotype representation, for instance, by extracting data from news in which positive attributes regarding immigrants are reported.

Thanks to the outcomes of the annotation procedure, we were able to perform a multilingual analysis and machine learning experiments in three languages on these texts, which have a Twitter conversation structure. Additionally, our results suggest that the degree of implicitness depends greatly on the dimension of contextuality in this domain.

Detecting stereotypes is an inherently difficult task, due to their overall low rate of occurrence on social media platforms, in part due to effective content moderation. Furthermore, the corpus and the set of analyses and experiments presented here were produced and performed in a multilingual context, and, as such, were conditioned by many language-specific characteristics and differences. These differences range from cultural, linguistic and geographical specificities, which may have impacted the data collection and annotation process, to size and distribution differences in the language-specific subsets, as well as the conversational context in which the messages were communicated.

Finally, despite their similar architectures, the three monolingual classification models were pre-trained not only on different languages and types of content but also with different hyperparameters and implementation details, which may have been sources of variations in their final predictive capabilities, even assuming the fine-tuning data were equivalent. Nevertheless, we believe that this work constitutes an important resource and basis for further work on multilingual stereotype detection in social media content.

Looking to the future, we may explore in more depth the language-specific differences and commonalities of stereotypes within the practical context of multilingual stereotypes classification across the three European languages which constitute our corpus. We will investigate whether language-universal patterns in stereotypes can be successfully exploited by state-of-the-art multilingual NLP architectures.

Thanks to the resources and framework elaborated in this study, it will also be possible to investigate more deeply the phenomenon of the dissemination of racial stereotypes through racial hoaxes from a computational perspective and in a multilingual context. Furthermore, these initial steps may prove essential for developing computational tools for the automatic detection and classification of racial stereotypes in real-life scenarios.

Acknowledgements This work was supported by the European project 'STERHEOTYPES—Studying European Racial Hoaxes and sterEOTYPES' funded by the Compagnia di San Paolo and VolksWagen Stiftung under the 'Challenges for Europe' call for Project (CUP: B99C20000640007). The research of CLiC was also partially funded by the 'FairTransNLP-Language' project (MCIN/AEI/10.13039/501100011033/ FEDER, UE, PID2021-124361OB-C33).

Author Contributions Based on the CRediT (Contributor Roles Taxonomy) author statement, Wolfgang S. Schmeisser-Nieto and Alessandra Teresa Cignarella performed conceptualization, methodology, formal analysis, investigation, resources, data curation, writing—original draft and visualization; Tom Bourgeade performed conceptualization, methodology, software, formal analysis, investigation, resources, data curation and writing—original draft; Simona Frenda performed conceptualization, methodology, formal analysis, investigation, resources and data curation; Alejandro Ariza-Casabona performed investigation and data curation; Giuseppe Corbelli and Paolo Cicirelli performed conceptualization and writing—original draft; Mario Laurent performed methodology, formal analysis, investigation, resources and data curation; Andrea Marra performed formal analysis; Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti and Mariona Taulé performed conceptualization, methodology, writing—review & editing, supervision and project administration; Marinella Paciello performed conceptualization, writing—original draft, supervision and project administration; and Francesca D'Errico performed conceptualization, methodology, writing—original draft, review editing, supervision, project administration and funding acquisition.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. The research of Farah Benamara is also partially supported by DesCartes: The National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

Availability of data and materials The StereoHoax corpus will be available upon request only for academic purposes. We have removed usernames to preserve anonymity.

Code availability The code used to extract our dataset and to run the models is available upon request.

Declarations

Conflict of interest The authors declare they have no financial interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication All the authors have given explicit consent to submit this work.

References

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: agency with competence and assertiveness-communion with warmth and morality. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.01810. http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01810/full

Allport, G. W. (1935). Attitudes. In *A handbook of social psychology* (pp. 798–844). CA Murchinson.

Allport, G. W., Clark, K., & Pettigrew, T. (1954). *The nature of prejudice*. Addison-Wesley Reading.

Arcuri, L. (2015). *Due pesi due misure. Come gli immigrati e gli italiani sono descritti dai media*. Giunti Editore.

Ariza-Casabona, A., Schmeisser-Nieto, W. S., Nofre, M., Taulé, M., Amigó, E., Chulvi, B., & Rosso, P. (2022). Overview of DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish. *Procesamiento del Lenguaje Natural, 69*, 217–228.

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education, 31*(2), 101–19.

Beukeboom, C. J., & Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research, 7*, 1–37.

Beukeboom, C. J., Finkenauer, C., & Wigboldus, D. H. J. (2010). The negation bias: When negations signal stereotypic expectancies. *Journal of Personality and Social Psychology, 99*(6), 978–99. https://doi.org/10.1037/a0020861

Biewald, L. (2020). Experiment tracking with weights and biases. https://www.wandb.com/, software available from wandb.com.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*, 4349–4357.

Bosco, C., Patti, V., Frenda, S., Cignarella, A. T., Paciello, M., & D'Errico, F. (2023). Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP. *Information Processing & Management,60*(1), 103118. https://doi.org/10.1016/j.ipm.2022.103118. https://linkinghub.elsevier.com/retrieve/pii/S0306457322002199.

Bourgeade, T., Cignarella, A. T., Frenda, S., Laurent, M., Schmeisser-Nieto, W. S., Benamara, F., Bosco, C., Moriceau, V., Patti, V., & Taulé, M. (2023) A multilingual dataset of racial stereotypes in social media conversational threads. In *Proceedings of the 17th conference of the European chapter of the association for computational linguistics (EACL 2023)*. In Press.

Bye, H. H. (2020). Intergroup relations during the refugee crisis: individual and cultural stereotypes and prejudices and their relationship with behavior toward asylum seekers. *Frontiers in Psychology, 11*, 612267. https://doi.org/10.3389/fpsyg.2020.612267. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.612267/full.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020) Spanish pre-trained Bert model and evaluation data. In *PML4DC at ICLR 2020*.

Cao, Y. T., Sotnikova, A., Daumé III, H., Rudinger, R., & Zou, L. (2022) Theory-grounded measurement of U.S. social stereotypes in English language models. In Carpuat, M., de Marneffe, M. C., Meza Ruiz, IV (Eds.), *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies. association for computational linguistics, Seattle, USA* (pp. 1276–1295). https://doi.org/10.18653/v1/2022.naacl-main.92. https://aclanthology.org/2022.naacl-main.92.

Card, D., Gross, J. H., Boydstun, A., & Smith N. A. (2016). Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1410–1420).

Cardellino, C. (2019). Spanish billion words corpus and embeddings. https://crscardellino.github.io/SBWCE/.

Castelfranchi, C. (2023). The nature of power and its complex dynamics. In *A Theory of Tutelary Relationships* (pp. 59–109). Springer.

Cerase, A., & Santoro, C. (2018). From racial hoaxes to media hypes: Fake news' real consequences. In P. Vasterman (Ed.), *rom media hype to twitter storm: new explosions and their impact on issues, crises, and public opinion* (pp. 333–354).

Cercas Curry, A., Abercrombie, G., & Rieser, V. (2021). ConvAbuse: data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 conference on empirical methods in natural language Processing. Association for computational linguistics, online* (pp. 7388–7403). https://doi.org/10.18653/v1/2021.emnlp-main.587. https://aclanthology.org/2021.emnlp-main.587.

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology, 104*(1), 103.

Chiril, P., Benamara, F., & Moriceau, V. (2021). "Be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the association for computational linguistics: EMNLP 2021. Association for computational linguistics* (pp. 2833–2844). https://doi.org/10.18653/v1/2021.findings-emnlp.242. https://aclanthology.org/2021.findings-emnlp.242.

Collins, K. A., & Clément, R. (2012). Language and prejudice: Direct and moderated effects. *Journal of Language and Social Psychology, 31*(4), 376–396. https://doi.org/10.1177/0261927X12446611

D'Errico, F., & Paciello, M. (2018). Online moral disengagement and hostile emotions in discussions on hosting immigrants. *Internet Research,28*(5), 1313–1335. https://doi.org/10.1108/IntR-03-2017-0119. https://www.emerald.com/insight/content/doi/10.1108/IntR-03-2017-0119/full/html.

D'Errico, F., Corbelli, G., Papapicco, C., & Paciello, M. (2022). How personal values count in misleading news sharing wh moral content. *Behavioral Sciences, 12*(9), 302. https://doi.org/10.3390/bs12090302. https://www.mdpi.com/2076-328X/12/9/302.

D'Errico, F., Papapicco, C., & Taulé Delor, M. (2022). 'Immigrants, hell on board': Stereotypes and prejudice emerging from racial hoaxes through a psycho-linguistic analysis. *Journal of Language and Discrimination*. https://doi.org/10.1558/jld.21228. https://journal.equinoxpub.com/JLD/article/view/21228.

D'Errico, F., & Poggi, I. (2014). Acidity. The hidden face of conflictual and stressful situations. *Cognitive Computation, 6*, 661–676.

Dev, S., Li, T., Phillips, J.M., & Srikumar, V. (2020). On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 7659–7666).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies (long and short papers)* (Vol. 1, pp 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423.

Esses, V. M., Medianu, S., & Lawson, A. S. (2013). Uncertainty, threat, and the role of the media in promoting the dehumanization of immigrants and refugees: Dehumanization of immigrants and refugees. *Journal of Social Issues, 69*(3), 518–536. https://doi.org/10.1111/josi.12027

Fersini, E., Nozza, D., & Rosso, P. (2020). AMI @ EVALITA2020: automatic misogyny identification. In: Basile, V., Croce, D., Maro, M. D., Lucia C. P. (Eds.), Proceedings of the 7th evaluation campaign of natural language processing and speech tools for Italian. Final workshop (EVALITA 2020), Online event, December 17th, 2020, CEUR workshop proceedings (Vol. 2765). CEUR-WS.org. http://ceur-ws.org/Vol-2765/paper161.pdf.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the Evalita 2018 task on automatic misogyny identification (AMI). *EVALITA Evaluation of NLP and Speech Tools for Italian, 12*, 59.

Fiske, S. (1998). Stereotyping, prejudice, and discrimination. In: Gilbert, D. T., Fiske, S. T. & Lindzey, G. (Eds.). *The handbook of social psychology* (pp. 357–411).

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005. https://linkinghub.elsevier.com/retrieve/pii/S1364661306003299.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67–73.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (Often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Fokkens, A., Ruigrok, N., Beukeboom, C, Sarah, G., & Van Atteveldt, W. (2018). Studying Muslim stereotyping through microportrait extraction. In *Proceedings of the 11th international conference on language resources and evaluation (LREC 2018)* (pp. 3734–3741).

Fraser, K. C., Kiritchenko, S., & Nejadgholi, I. (2022a). Computational modeling of stereotype content in text. *Frontiers in Artificial Intelligence*, *5*.

Fraser, K. C., Kiritchenko, S., & Nejadgholi, I. (2022b). Extracting age-related stereotypes from social media texts. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., & Piperidis, S. (Eds.), *Proceedings* *of the 13th language resources and evaluation conference* (pp. 3183–3194). European Language Resources Association. https://aclanthology.org/2022.lrec-1.341.

Fraser, K. C., Kiritchenko, S., & Nejadgholi, I. (2024). How does stereotype content differ across data sources? In *Proceedings of the 13th joint conference on lexical and computational semantics* (*SEM 2024) (pp. 18–34).

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*, 252–64.

Herold, B., Waller, J., & Kushalnagar, R. (2022). Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In Ebling, S., Prud'hommeaux, E., & Vaidyanathan, P. (Eds.), 9th Workshop On Speech And Language Processing For Assistive Technologies (slpat-2022) (pp. 58–65). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.slpat-1.8. https://aclanthology.org/2022.slpat-1.8.

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The abc of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110*(5), 675.

Lee, T. L., & Fiske, S. T. (2006). Not an outgroup, not yet an ingroup: immigrants in the stereotype content model. *International Journal of Intercultural Relations, 30*(6), 751–768. https://doi.org/10.1016/j.ijintrel.2006.06.005. https://linkinghub.elsevier.com/retrieve/pii/S0147176706000526.

Maass, A., Karasawa, M., Politi, F., & Suga, S. (2006). Do verbs and adjectives play different roles in different cultures? A cross-linguistic analysis of person representation. *Journal of Personality & Social Psychology, 90*(5), 734–750.

Markov, I., & Daelemans, W. (2022). The role of context in detecting the target of hate speech. In *Proceedings of the 3rd workshop on threat, aggression and cyberbullying (TRAC 2022)* (pp. 37–42). Association for Computational Linguistics. https://aclanthology.org/2022.trac-1.5.

Marshall, S. R., & Shapiro, J. R. (2018). When "scurry'' vs. "hurry'' makes the difference: Vermin metaphors, disgust, and anti-immigrant attitudes. *Journal of Social Issues, 74*(4), 774–789.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., & Sagot, B. (2020). CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7203–7219). Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.645. https://aclanthology.org/2020.acl-main.645.

Menini, S., Aprosio, A. P., & Tonelli, S. (2021). Abuse is contextual, What about NLP? The role of context in abusive language annotation and detection. https://doi.org/10.48550/arXiv.2103.14916.

Mina, M., Falcão, J., & Gonzalez-Agirre, A. (2024). Exploring the relationship between intrinsic stigma in masked language models and training data using the stereotype content model. In Kokkinakis, D., Fraser, K. C., Themistocleous, C. K, Fors, K. L., Tsanas, A., & Ohman, F. (Eds.), *Proceedings of the 5th workshop on resources and processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024* (pp. 54–67). ELRA and ICCL, Torino, Italia. https://aclanthology.org/2024.rapid-1.7.

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology, 51*(1), 178–196.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). Toxicity detection: Does context really matter? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4296–4305). Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.396. https://aclanthology.org/2020.acl-main.396.

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in western Europe. *European Journal of Social Psychology, 25*(1), 57–75. https://doi.org/10.1002/ejsp.2420250106

Poggi, I., & D'Errico, F. (2010). Dominance signals in debates. In Salah, A. A., Gevers, T., Sebe, N., et al. (Eds.), Human behavior understanding (Vol. 6219, pp. 163–174). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14715-9_16.

Poggi, I., D'Errico, F., & Vincze, L. (2011a). Agreement and its multimodal communication in debates: A qualitative analysis. *Cognitive Computation, 3*, 466–479.

Poggi, I., D'Errico, F., & Vincze, L. (2011b). Discrediting moves in political debate. In *Proceedings of second international workshop on user models for motivational systems: The affective and the rational routes to persuasion (UMMS 2011)(Girona) Springer LNCS* (pp. 84–99).

Rodríguez-Sánchez, F., de Albornoz, J. C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., & Donoso,T. (2021) Overview of EXIST 2021: Sexism identification in social networks. *Procesamiento del Lenguaje Natural*, *67*, 195–207. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

Rodríguez-Sánchez, F., de Albornoz, J. C., Plaza, L., Mendieta-Aragn, A., Marco-Remn, G., Makeienko, M., Plaza, M., Gonzalo, J., Spina, J., Rosso, P. (2022) Overview of EXIST 2022: Sexism identification in social networks. *Procesamiento del Lenguaje Natural*, *69*, 229–240. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

Russell, K. K. (1998). *The color of crime: Racial hoaxes, white fear, black protectionism, police harassment, and other macroaggressions*. New York University Press.

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

Sánchez-Junquera, J. J., Rosso, P., Montes, M., Chulvi, B., & others. (2021). Masking and BERT-based models for stereotype identication. *Procesamiento del Lenguaje Natural, 67*, 83–94.

Sanguinetti, M., Comandini, G., di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti,V., & Russo, I. (2020) Haspeede 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. In: Basile, V., Croce, D., Di Maro, M., & Passaro, L. C. (Eds.), *Proceedings of the 7th evaluation campaign of natural language processing and speech tools for Italian. Final workshop (EVALITA 2020). CEUR workshop proceedings (CEUR-WS.org), conference, 17 Dec 2020* (Vol. 2765).

Sap, M., Gabriel, S., Qin, L., Jurafsky, D.,Smith, N. A., & Choi, Y. (2020) Social bias frames: Reasoning about social and power implications of language. arXiv:1911.03891.

Schmeisser-Nieto, W. S., Nofre, M., & Taulé, M. (2022). Criteria for the annotation of implicit stereotypes. In *Proceedings of the 13th language resources and evaluation conference (LREC 2022)* (pp. 753–762).

Suárez, P. J. O., Sagot, B., & Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th workshop on the challenges in the management of large corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache. https://doi.org/10.14618/IDS-PUB-9021. https://hal.inria.fr/hal-02148693.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2214–2218). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Ungless, E. L., Rafferty, A., Nag, H., & Ross, B. (2022). A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the 5th workshop on natural language processing and computational social science (NLP+CSS)* (pp. 207–217).

Utych, S. M. (2018). How dehumanization influences attitudes toward immigrants. *Political Research Quarterly, 71*(2), 440–452.

Vaes, J., Latrofa, M., Suitner, C., & Arcuri, L. (2017). They are all armed and dangerous! *Journal of Media Psychology*, *31*.

Vargas, F., Carvalho, I., Hürriyetoğlu, A., Pardo, T., & Benevenuto, F. (2023). Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In Mitkov, R., Angelova, G. (Eds.), *Proceedings of the 14th international conference on recent advances in natural language processing* (pp. 1187–1196). INCOMA Ltd. https://aclanthology.org/2023.ranlp-1.126.

Wardle, Derakhshan. H. (2018). Thinking about 'information disorder': Formats of misinformation, disinformation, and mal-information. In *Journalism, 'fake news' & disinformation*. Unesco.

Wenzek, G., Lachaux, M. A., Conneau A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2020) CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th language resources and evaluation conference* (pp. 4003–4012). European Language Resources Association. https://aclanthology.org/2020.lrec-1.494.

Wright, C., Brinklow-Vaughn, R., Johannes, K., & Rodriguez, F. (2021). Media portrayals of immigration and refugees in hard and fake news and their impact on consumer attitudes. *Howard Journal of Communications, 32*(4), 331–351. https://doi.org/10.1080/10646175.2020.1810180

Ybarra, O. J., & Stephan, W. G. (1994). Perceived threat as a predictor of stereotypes and prejudice: Americans' reactions to Mexican immigrants. *Boletin de Psicologia*, *42*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Wolfgang S. Schmeisser-Nieto[1] · **Alessandra Teresa Cignarella**[2,3] · **Tom Bourgeade**[4,5] · **Simona Frenda**[2,6] · **Alejandro Ariza-Casabona**[1] · **Mario Laurent**[4] · **Paolo Giovanni Cicirelli**[8] · **Andrea Marra**[2] · **Giuseppe Corbelli**[9] · **Farah Benamara**[4,9] · **Cristina Bosco**[2] · **Véronique Moriceau**[4] · **Marinella Paciello**[7] · **Viviana Patti**[2] · **Mariona Taulé**[1] · **Francesca D'Errico**[8]

✉ Wolfgang S. Schmeisser-Nieto
wolfgang.schmeisser@ub.edu

Alessandra Teresa Cignarella
alessandrateresa.cignarella@ugent.be

Tom Bourgeade
tom.bourgeade@loria.fr

Simona Frenda
s.frenda@hw.ac.uk

Alejandro Ariza-Casabona
alejandro.ariza14@ub.edu

Mario Laurent
mario.laurent@irit.fr

Paolo Giovanni Cicirelli
paolo.cicirelli@uniba.it

Andrea Marra
andrea.marra@unito.it

Giuseppe Corbelli
g.corbelli@students.uninettunouniversity.net

Farah Benamara
farah.benamara@irit.fr

Cristina Bosco
cristina.bosco@unito.it

Véronique Moriceau
veronique.moriceau@irit.fr

Marinella Paciello
marinella.paciello@uninettunouniversity.net

Viviana Patti
viviana.patti@unito.it

Mariona Taulé
mtaule@ub.edu

Francesca D'Errico
francesca.derrico@uniba.it

[1] Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona, Barcelona, Spain

[2] Dipartimento di Informatica, Università di Torino, Turin, Italy

[3] Present Address: LT3, Ghent University, Ghent, Belgium

[4] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France

[5] Present Address: CNRS, Inria, LORIA, Université de Lorraine, Nancy, France

[6] Present Address: Interaction Lab, Heriot-Watt University, Edinburgh, Scotland

[7] Facoltà di Psicologia, Università Telematica Internazionale UniNettuno, Rome, Italia

[8] Dipartimento ForPsiCom, Formazione, Psicologia, Comunicazione, Università di Bari “Aldo Moro”, Bari, Italy

[9] IPAL, CNRS-NUS-A*STAR, Singapore, Singapore