# A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang S. Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, Mariona Taulé

# A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads

**Tom Bourgeade**[1,*], **Alessandra Teresa Cignarella**[2,*], **Simona Frenda**[2,*], **Mario Laurent**[1], **Wolfgang S. Schmeisser-Nieto**[3], **Farah Benamara**[1,4], **Cristina Bosco**[2], **Véronique Moriceau**[1], **Viviana Patti**[2] and **Mariona Taulé**[3]
1. IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
2. Dipartimento di Informatica, Università degli Studi di Torino, Italy
3. CLiC research group, UBICS, Universitat de Barcelona, Spain
4. IPAL, CNRS-NUS-ASTAR, Singapore

**Warning:** *This paper contains examples of potentially offensive content.*

## Abstract

In this paper, we focus on the topics of misinformation and racial hoaxes from a perspective derived from both social psychology and computational linguistics. In particular, we consider the specific case of anti-immigrant feeling as a first case study for addressing racial stereotypes. We describe the first corpus-based study for multilingual racial stereotype identification in social media conversational threads. Our contributions are: (i) a multilingual corpus of racial hoaxes, (ii) a set of common guidelines for the annotation of racial stereotypes in social media texts, and a multi-layered, fine-grained scheme, psychologically grounded on the work by Fiske et al., including not only stereotype presence, but also contextuality, implicitness, and forms of discredit, (iii) a multilingual dataset in Italian, Spanish, and French annotated following the aforementioned guidelines, and cross-lingual comparative analyses taking into account racial hoaxes and stereotypes in online discussions. The analysis and results show the usefulness of our methodology and resources, shedding light on how racial hoaxes are spread, and enable the identification of negative stereotypes that reinforce them.

## 1 Introduction

Racial Hoaxes (RHs) are "a communicative act oriented to spread fallacious information against a social group" (Russell, 1998). As social media have become a dominant means of communication, investigating them is crucial for tackling the spread of RHs. We approach this task combining psychological and computational linguistics methods with a multilingual, cross-cultural perspective (Italian, Spanish, and French).

In particular, RHs can contribute to the diffusion of stereotypes about people belonging to the *outgroup*, i.e., a social group with features that differ from the *ingroup* (Rooduijn et al., 2021) and are, thus, more vulnerable. Even common, naive users are as likely to become spreaders of RHs as malicious users (Papapicco et al., 2022). In this paper, we cover a specific theme: anti-immigrant stereotypes. The discursive construction of immigrants and refugees in user interaction on social media has been studied by Ekman (2019), who has shown how racial expressions and overt racism are becoming increasingly normalized, thus leading to prejudices and racial stereotypes and, eventually, even harmful acts.

Overall, the attention to these topics is relatively new in the NLP community, and thus, there is still a meaningful lack of annotated resources for the development of automatic tools to detect stereotypes and related phenomena. Among the few research contributions in this direction, Sanguinetti et al. (2020) organized the second edition of HaSpeeDe at EVALITA 2020, asking participants to automatically detect hate speech and stereotypes in Italian tweets and headlines. Similarly, in the DETESTS shared task at IberLEF 2022, Ariza-Casabona et al. (2022) proposed a 10-label classification task for the identification of stereotypes in Spanish; and finally, during IROSTEREO at PAN/CLEF 2022, Ortega-Bueno et al. (2022) proposed an author profiling task regarding stereotype spreaders and studied the link with irony in English. Recently, for French, Chiril et al. (2021) investigated how to improve gender hate speech classification by leveraging stereotype detection based on multitask architectures.

* The first three authors contributed equally.



*Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696
May 2-6, 2023 

However, such related works only focus on monolingual contents, without considering multilingual settings from which cross-cultural differences and similarities in the expression of stereotypes can emerge. Furthermore, most of the related work limits the scope of investigation to the mere presence/absence of stereotypes in a single text, without diving into the finer-grained features that arise from psychological studies (Allport et al., 1954; Fiske et al., 2007; Cuddy et al., 2008), and without taking into account their propagation in social media conversational threads. Considering the gaps in current related work, we propose a cross-cultural, and multilingual perspective for studying racial hoaxes and stereotypes. In this work, our original contributions are:

- A Multilingual Racial Hoaxes Corpus that was manually created, extracting fake news about migration and racial content from fact-checking web sites. The list of hoaxes has been employed as the core knowledge-base for extracting texts from social media that spread RHs and the *reactions* to them.[1]
- A methodology that makes it possible to collect a full conversational thread, with replies and comments that are written under the post spreading the main racial hoax.
- A multi-layered annotation scheme for the annotation of racial stereotypes in social media texts, which allows us to study how the presence of a racial hoax interacts with the surrounding textual context. The scheme, based on psychological work by Fiske (1998), includes four layers: (a) stereotype presence, (b) contextuality, (c) implicitness and (d) forms of discredit.
- A multilingual dataset annotated according to this scheme. For this first study, we chose to retrieve data in languages that are spoken in three countries on the maritime coast of the Mediterranean basin, where migration is widespread and has been made a particular issue in local politics: Italy, Spain and France.[2]
- Qualitative and quantitative analyses from a comparative perspective of the three language subsets, focusing in particular on the interactions between the topics of RHs, stereotypes and discredit in conversations.

## 2 Related Work

### 2.1 RH and stereotypes in Psychology

Hoaxes are a form of 'misinformation' that aims to disseminate false information with the intention of making it viral in social media (Wardle and Derakhshan, 2018). In particular, 'Racial Hoaxes' are fallacious discursive acts that contribute to the spread of information against a social group because of race, religion or origin, such as 'immigrants' (Cerase and Santoro, 2018).

From a psychological point of view, RHs have become an important object of study since, firstly, they help to spread misinformation by attacking, discrediting and damaging immigrants' image; secondly, they can increase the formation of people's prejudices and stereotypes towards the *outgroup* (Fiske, 1998). In fact, while the stereotype is the cognitive nucleus of prejudice, which contains a set of beliefs and social images; prejudice is a preconceived attitude that is based on common voices and opinions. RHs, therefore, appear to install a stereotype facilitating a categorization in which there is a generalization through a label referring to an entire group, e.g., 'all immigrants are thieves' (Allport et al., 1954).

The manifestations of stereotypes can range from a more explicit to a more implicit expression. It is possible, in fact, to distinguish an EXPLICIT stereotype content when identifying a direct association between immigrants and a particular quality, e.g., 'immigrants bring us diseases' (Fiske and Taylor, 2013). IMPLICIT stereotypes can be expressed through evaluative utterances and figures of speech such as metaphors, humor, and irony. For instance, Schmeisser-Nieto et al. (2022) present criteria to identify and annotate implicit stereotypes focusing on immigration.

### 2.2 Stereotypes in Computational Linguistics

The computational linguistics community has only recently focused on modeling stereotypes in order to automatically recognize them, e.g., within political debates (Sánchez-Junquera et al., 2021a) or

[1] By 'reactions' we refer to replies and comments to the main thread that is spreading a racial hoax.

[1] To guarantee anonymity and protect the privacy of Twitter users, throughout this paper, instead of using direct quotations from the tweets, we only provide their English translations and/or adaptations.

[2] The annotated dataset will be available for research purposes upon request, together with the complete set of annotation guidelines.

social media (Sanguinetti et al., 2020; Chiril et al., 2021), but without considering the conversational threads in which they occur, nor their reinforcement or confirmation through RHs.

Recently, Sánchez-Junquera et al. (2021a) proposed a taxonomy of stereotypes about immigrants and approached the problem of the automatic classification of stereotypes in Spanish by focusing on the narrative *frames* that spread the stereotypes. Similarly, Fokkens et al. (2018) approached stereotype detection by extracting the *microportraits* and Card et al. (2016) by extracting *stories* about individuals from text. Beukeboom and Burgers (2019) propose a framework which looks at how stereotypes are shared through language: bias in labels and bias in the description of characteristics and behaviors.

Fraser et al. (2022) rely on the Stereotype Content Model (SCM) and present a computational method to mine large datasets and then map sentences to the two-dimensional plane of perceived *warmth* and *competence* (Fiske et al., 2007). Other common computational approaches in NLP mainly focused on measuring and quantifying social bias towards different groups, especially using techniques of word representation, such as word embedding (Bolukbasi et al., 2016), transformers (Card et al., 2016), techniques of natural language inference (Dev et al., 2020) and masking BERT for racial stereotype detection (Sánchez-Junquera et al., 2021b). In this context, this multidisciplinary study on the stereotypes related to RHs from a multilingual, cross-cultural perspective represents an interesting, novel opportunity to understand the expression, perception, and reinforcement of stereotypes, stemming from RHs, against immigrants in conversations on Twitter.

## 3 From Racial Hoaxes to Reactions

In order to collect reactions to racial hoaxes on social media, we first created the Multilingual Racial Hoaxes Corpus (MRHC), a list of 239 RHs in three languages: Italian, Spanish, and French. Given the difficulty of spotting them automatically, we collected the entries of the MRHC manually.

Depending on the language, different fact-checking websites or newspapers commenting on hoaxes were used as a source for manually extracting the MRHC between 2019 and 2021. For instance, for Italian we used the debunking sites `bufale.net` and `butac.it`; for Spanish `maldita.es` and `newtral.es`; and finally for French `factuel.afp.com` and `lemonde.fr/les-decodeurs`.

### 3.1 Topics of the MRHC

Inspired by the taxonomy of stereotypes proposed in Sánchez-Junquera et al. (2021a); Ariza-Casabona et al. (2022), we defined five macro categories of topics, in which immigrants are perceived as threat by the society.

Table 1 contains some examples for each topic: **(a) Security** for events related to citizen safety, such as murder, sexual assault, fights, terrorist attacks, theft, and public disorder; **(b) Public Health** related to health issues that may potentially affect the population, mainly infectious diseases (e.g., COVID-19); **(c) Migration Control** covers migratory flows, arrivals, disembarkation, border control and the regulation of immigration; **(d) Benefits** describe situations in which the outgroup (immigrants) receives more help, social assistance and welfare benefits than the ingroup; **(e) Religion** covers religious and cultural differences of the out group that threaten the traditions of the ingroup (even though terrorism and religion are closely associated in RH, the former category has been considered under the security topic), and finally, **(f) Others** includes RHs about other topics not included in the previous categories.

In terms of a cross-cultural analysis, we observed variations among the different types of RHs. As shown in Table 2, the most common topic of RHs in Italian is related to Security, accounting for 58.76% of the total, while in Spanish and French, RHs are related to Benefits, accounting for 29.16% and 50% respectively. Another relevant result is that the topic Religion has no representation in the Italian subset, which is also the case of Public Health in the French subset.

### 3.2 Reactions to RHs

We started the collection procedure by retrieving texts from Twitter that contained one of the RHs from the MRHC, or texts that presented a high similarity to one of those. We searched for texts containing the same URL as the RH, or same title of news of the RH on the debunking sites, or even keywords extracted from the textual body of RH by using the `Twitter APIs v2` for Academia.[3] In

[3] `https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries/v2`

| Example | Topic |
|---|---|
| Immigrants out of control: they flee and injure an officer | Security |
| Migrant with Covid repatriated. And now 100 agents are in quarantine | Public Health |
| The electoral roll increases because the Government nationalizes 200,000 "illegals" | Migration Control |
| A foreign minor, 4,700€ per month, your grandmother, 426€ pension per month | Benefits |
| In Aubervilliers, the sheep ready to be slaughtered for #Eid on their way to the butcher. Mind boggling! #Ramadam | Religion |

Table 1: Examples of different topics of RHs. All tweets were originally written either in Italian, Spanish or French. They have been translated to English and adapted to ensure anonymity and guarantee privacy to users.

| 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 |

Table 2: Percentages of Types of RHs in the three language subsets.

Figure 1 we show the full pipeline employed for the collection of "reactions to racial hoaxes".

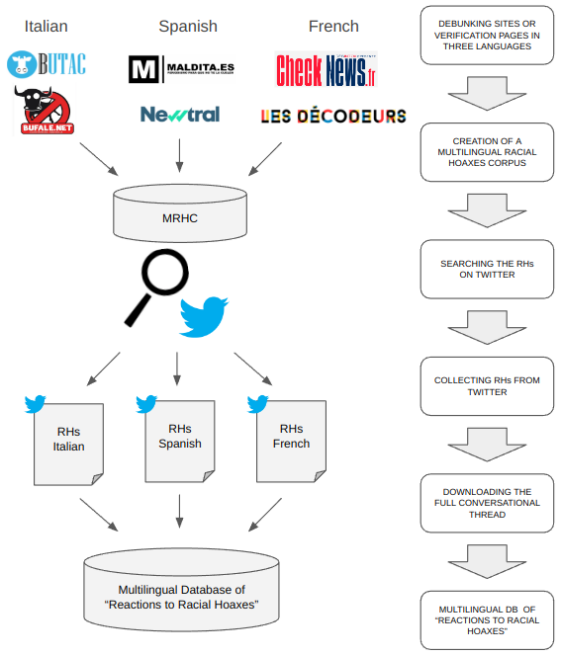


Figure 1: Pipeline for the creation of the Multilingual Racial Hoaxes Corpus (MRHC) and reactions to them.

As can be seen from the picture above, when a racial hoax from the MRHC was found on Twitter, we referred to it as the 'Conversational Head', because it was the first text in the conversational thread. Then, for each language, we retrieved all the conversational heads and, in order to study the conversational context, we further collected all the direct replies, and the replies-to-replies.

After the collection and cleaning of data, we obtained a total of 2,850 unique tweets stemming from Conversational Heads for Italian, 4,751 tweets for Spanish, and 9,305 tweets for French. In Table 3 we display the information on the three subsets of the multilingual dataset. We show the number of the original RHs that we searched for on Twitter and from which we were able to extract the Conversational Heads. In the other columns, we display the number of direct replies, the number of replies-to-replies, and the total of reactions (tweets). In many cases, we had to discard the original RH because it did not originate a conversational thread on Twitter but rather contained just images, videos or recording from other platforms that have not been commented on Twitter with textual content (see the difference between the numbers in the first two columns of Table 3).

## 4 Annotating Reactions to Racial Hoaxes

### 4.1 A Multi-layered Annotation Scheme

The annotation scheme designed for the multilingual dataset is inspired by studies regarding stereotypes in the psychological and linguistic literature (Fiske et al., 2007; Cuddy et al., 2008; Sánchez-Junquera et al., 2021a). The outcome of such research is a scheme that consists of four layers, organized in two levels:

1. The first level refers to the presence of a **racial stereotype** as a binary category (*yes*/*no*).

2. The second level can be annotated only if the

| Lang. | Original RHs | RHs found on Twitter | Conversational Heads | Direct Replies | Replies to Replies | Total of Reactions |
|---|---|---|---|---|---|---|
| Italian | 97 | 50 | 273 | 597 | 2,253 | 2,850 |
| Spanish | 72 | 24 | 353 | 85 | 4,313 | 4,751 |
| French | 70 | 36 | 36 | 3,927 | 5,378 | 9,305 |

Table 3: Number of RHs and details about conversational threads.

precedent level is annotated as *yes*, and it includes three categories:

(a) **Contextuality**. It encodes whether, in order to understand the meaning of the racial stereotype expressed, you need to look through the context (such as Twitter thread, the RH that triggered the conversation, URLs and images). It is annotated as a binary category (*yes*/*no*).

(b) **Implicitness**. It encodes whether the stereotype is expressed explicitly in the message (i.e., a clear span of text where lexical items can be selected) or whether at least one inference needs to be made for the stereotype to be understood). It is annotated as a binary category (*explicit*/*implicit*).

(c) **Forms of Discredit**. It encodes the precise form in which the text spreads a racial or anti-migration stereotype, attributing a type of behavior to the discriminated target. The values that can be applied are six: Affective Competence (AC), Attack to Benevolence (B), Competence (C), Dominance Down (DD), Dominance Up (DU) and Physical (P).

These six categories inspired by the Stereotype Content Model proposed by Fiske (1998), can in turn be encompassed in two: COMPETENCE (including C, DD, P) and WARMTH (including AC, B, DU). In the SCM, these macro-categories are respectively referred to as "*agency*" and '*communion*". For instance, Cuddy et al. (2008) show how, depending on the emotion that is elicited primarily by the form of discredit, different ways of sorting and grouping could be admissible. Furthermore, they underline that the main dimensions of COMPETENCE and WARMTH can be seen as a two-dimensional array for sorting groups. This is an ideal solution that includes at least four clusters which significantly differ regarding warmth and competence.

This motivates our strategy in which Competence (C) is grouped with Physical (P) (both forms of discredit with HIGH COMPETENCE), and Attack to Benevolence (B) with Dominance Up (DU) (both forms of discredit with LOW WARMTH), resulting in the following four clusters for forms of discredit: C+P, DD, B+DU, AC.[4]

### 4.2 Annotation and Agreement

The data were entirely annotated on locally adapted versions of the LabelStudio[5] open source platform, in which the questions and labels of the annotation scheme were translated into all the three languages.

The Italian portion of the dataset was annotated by two trained native speakers. Concerning the main dimension of stereotype, they obtained an inter-annotator agreement (IAA) of $\kappa = 0.48$, as calculated by Cohen's kappa coefficient (moderate). The remaining disagreement was solved by a third expert. The Spanish subset was annotated by three annotators, two of whom are Linguistics students trained for the task, along with a researcher. The IAA was calculated by Fleiss' Kappa coefficient, resulting in $\kappa = 0.76$. The French subset was annotated by a total of four annotators: an expert and three Linguistics students. Due to the larger quantity of data to annotate, most of the subset was annotated separately by two annotators (two sets of $\sim 4,250$ tweets). The rest was annotated in three sets, each by two annotators, at different stages of the annotation process, to ensure no degradation in IAA was occurring. The Cohen's Kappa for the French stereotype annotations is $\kappa = 0.73$.

Comparing the scores in the three subsets, it can be noticed that in Italian the IAA is lower with respect to those obtained in French and Spanish. Our hypothesis to explain this is linked to the fact that, in Italian conversational threads, the discussions among users tends to shift quickly to other sub-

[4]Please note that the dataset has been annotated according to the six forms of discredits and that this grouping has been designed with a computational perspective in mind.

[5]https://labelstud.io/

jects that are unrelated to RHs. We think that this conversational drift in a large number of tweets created doubt among the annotators and lowered the overall IAA.

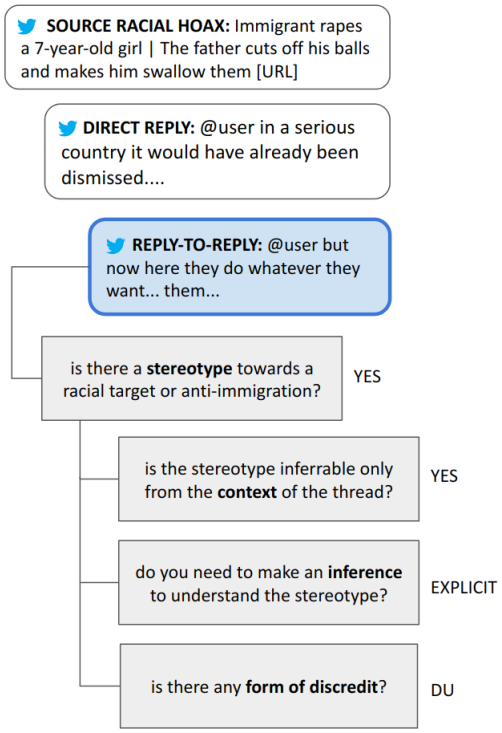


Figure 2: Triplet of tweets from a conversational thread, with the decision tree of the annotation scheme.

We conclude this section with a commented example. Figure 2 shows a Twitter conversational thread and the application of the annotation scheme on it. By looking at the third tweet of the triplet –in the blue box– it can be observed how the user reinforces the stereotypical distinction between "us" and "them", which highlights the concept of the *ingroup* as different from the *outgroup*. The anaphorically referenced "them" is the group (*outgroup*) to which the immigrant cited in the SOURCE RACIAL HOAX belongs, and for this reason the text has been annotated as containing a **racial stereotype**.

In order to grasp the presence of the stereotype and understand its content, the annotator also had to read the previous textual context (DIRECT REPLY and SOURCE RACIAL HOAX), so the dimension of **contextuality** was annotated as positive. As for the implicitness dimension, the tweet clearly states that "they do whatever they want", and because this sentence is a clear lexical expression of generalization, the stereotype is annotated as **explicit**. Finally, according to what the user wrote, the immigrant exercises a sort of forceful dominion and displays aggressive behavior, breaking the law. For this reason, the text was annotated as containing the form of discredit labelled **Dominance Up**.

## 5 Cultural and Linguistic Analyses

In this section, we describe the comparative analyses we performed to extract analogies and differences in the expression of stereotypes and the forms of discredit in the reactions to RHs among the three subsets.

### 5.1 Quantitative Results

In Table 4 we report the distribution and percentage of each annotated dimension. As can be seen, in the Italian and French data, stereotypes are found more rarely than in the Spanish subset, which contains about 30% of stereotypes. Another commonality between the Italian and French subsets is the distribution of contextuality and implicitness. In contrast, the Spanish subset contains a higher percentage of explicit stereotypes. 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 benefits by governments (DD), as well as criminality, illegality and fear of invasion (B+DU). In Italian, this last form of discredit is present with a higher percentage, followed by discredit regarding the competences of immigrants and their physical attributes (C+P).

In our dataset, the number of tweets containing stereotype is lower than in other datasets labelling the presence of this phenomenon (Sanguinetti et al., 2020; Ortega-Bueno et al., 2022). Rather than a purposely balanced dataset created in the context of shared tasks, our multilingual dataset is a reflection of users' reactions to RHs in social media.

### 5.2 Stereotype, Discredit, and Types of RHs

In this section, we report some observations regarding the reactions to RHs retrieved from Twitter in the three languages. For Italian, we were able to retrieve a total of 67 RHs on Twitter from the original 97 taken from fact-checking websites (see Table 2). However, after the annotation process and discussion, the gold dataset contains reactions to only 50 RHs. Those RHs that foster the

| Language | Tweets | Stereotype | | Contextual | | Implicitness | | Forms of Discredit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | *Agency* | | *Communion* | |
| | | *yes* | *no* | *yes* | *no* | *explicit* | *implicit* | *C+P* | *DD* | *B+DU* | *AC* |
| Italian | 2,850 | 234<br>8.21% | 2,616<br>91.79% | 177<br>75.64% | 57<br>25.36% | 95<br>40.60% | 138<br>59.40% | 71<br>23.75% | 40<br>13.38% | 176<br>58.86% | 12<br>4.01% |
| Spanish | 4,751 | 1,449<br>30.50% | 3,302<br>69.50% | 549<br>37.89% | 900<br>62.11% | 1,344<br>92.75% | 105<br>7.25% | 23<br>1.74% | 761<br>57.48% | 421<br>31.79% | 119<br>8.99% |
| French | 9,305 | 1,093<br>11.75% | 8,211<br>88.25% | 818<br>74.84% | 275<br>25.16% | 114<br>10.43% | 979<br>89.57% | 43<br>3.76% | 609<br>53.23% | 395<br>34.53% | 97<br>8.48% |

Table 4: Number of texts and label distribution for the categories annotated in the three language subsets. The numbers in the last four columns do not sum up to the total of the tweets containing stereotype. Indeed, discredit could be annotated with more than one label per tweet, and tweets could therefore be counted more than once.

| | Language | Benefit | Security | Migration Control | Public Health | Religion | Others |
|---|---|---|---|---|---|---|---|
| **Stereotype** | Italian | 0.21% | 51.69% | - | 48.09% | - | - |
| | Spanish | 38.79% | 20.01% | 10.97% | 0.07% | 30.16% | - |
| | French | 70.91% | 9.70% | 10.43% | - | 8.97% | - |

Table 5: Percentage co-occurrence of the presence of racial stereotypes and topic of the RH originally spread.

most stereotyped conversations are mainly nine, describing immigrants as threats to public health and security (see Table 5), as shown in the following examples:

(1) Coronavirus spreads, Government goes to secretly take illegal immigrants in Africa

(2) He kills an old Jewish woman at the cry of Allah Akbar. Acquitted because he was drugged.

The special attention paid to these two topics is also evident in the analysis of the most used hashtags in the tweets labelled with the presence of negative stereotypes, such as: #CrimesImmigrants, #SALVINI, #PD, #M5S, #hospitality. By using these hashtags, the users discuss the adopted policies of hospitality and control of immigration of various political parties (#SALVINI, #PD, #M5S), or depict immigrants as criminals (#CrimesImmigrants). The tweets containing these hashtags tend to be labelled with the B+DU form of discredit.

For Spanish, we were able to retrieve 24 RHs on Twitter, out of the 72 RHs originally collected from the fact-checking websites. The most prevalent topic within the Spanish context is related to benefits and the "illegality" of the immigrant. Those topics are associated directly to the forms of discredit DD and B+DU. These topics are also reflected in the use of hashtags such as: #StopIllegalImmigration or #Pensions.

Regarding French, from the 70 RHs identified at the start on the fact-checking website, we extracted 36 instances published on Twitter. As mentioned in Section 3.2, in some cases, we discarded the original RH because it did not originate a conversational Twitter thread or only contained images and videos without textual content. This was common in all the three languages considered.

Overall, French RHs had two common themes: attributing the role of victims to the representatives of Western civilization and the role of perpetrator to immigrants, as in Example (3) below; and pointing the finger at political decisions, real or fantasized, which would favor migrant populations at the expense of the "good French" such as farmers and students, as in Example (4).

(3) Immigrants burn down a refugee center because there's not enough Nutella: [URL]

(4) An immigrant who has never paid taxes in France receives 820 euros per month from the state, in the meantime some farmers get only 360 euros, how do you expect French people not to be angry?

The tweets similar to Example (3) are mainly associated with reactions containing B+DU types of discredit (around 35% of the total) while those similar to Example (4) are linked to DD (∼53%).

### 5.3 Contextuality and Implicitness

Focusing on the textual context, we analyzed how the stereotypes are propagated from the starting

point of the conversations throughout the thread, and if the context is needed to infer implicit forms of stereotypes in the three languages.

In the Italian subset, the majority of the tweets (87%) that are conversational heads (see Table 3) contain negative stereotypes against immigrants. However, even though a conversational head is deemed stereotypical, it is not correct to assume that all the tweets within its thread also contain stereotypes. Indeed, only about 17% of direct replies contain stereotypes and only 6% of the remaining threads are labelled with the presence of stereotypes. This is due mainly to two factors: 1) the tweets spreading fake news or offensiveness tend to be deleted by Twitter; 2) some of the tweets in the conversational threads tend to unveil the inaccuracy of the hoax.

Similarly to what happens in the Italian conversations, in the French subset, all conversational heads contain stereotypes, while 14% of direct replies and 10% of replies-to-replies contain stereotypes. Indeed, the fact that the RHs were debunked by fact-checking websites leads many comments to be criticisms of the conversational head, and this phenomenon is accentuated even more when the RH is shared by accounts with many followers. For the Spanish dataset, only 54% of conversational heads contain stereotypes, with the vast majority of stereotypes contained in replies, accounting for 90% of them.

| | | **Implicitness** |
|---|---|---|
| | Language | $\chi^2$ |
| **Contextuality** | Italian | 45.954 |
| | Spanish | 41.169 |
| | French | 11.419 |

Table 6: Association between contextuality and implecitenss. The $\chi^2$ tests are statically significant at $p < 0.001$ for the three languages.

The results reported in Table 6 show a statically significant association between the dimensions of implicitness and contextuality. As defined in Section 4, annotators labeled the necessity to use the context to understand the message or infer the presence of stereotypes. As expected, in the three datasets, the inference of stereotypes is especially facilitated by access to the textual context.

### 5.4 Lexical Analysis

To better understand the similarities and differences at the linguistic and cultural level between the languages, we performed a linguistic analysis, looking at the discriminative lexica used in texts containing stereotypes and labeled with specific forms of discredit. In particular, for all datasets, we listed: the most relevant n-grams[6] of the data annotated with stereo = *yes* (comparing them with the n-grams of the data annotated with stereo = *no*), and the most relevant n-grams from the data annotated with the four forms of discredit.

By looking at the resulting lists of words, we noticed that, in **Italian**, the words extracted from texts that do not contain stereotypes are related to the emotional sphere ("feeling", "feel ashamed", "hope"), in contrast to those extracted from texts containing stereotypes, which are related mainly to the negative actions of immigrants ("immigrant rape", "kill", "spit"). Regarding the various forms of discredit, we observed interesting differences. In general, words such as "invasion", "occupation" and "commanding" or expressions like "walk in underwear" or "laugh in court" are typical in texts annotated with the labels grouped under *communion*. In contrast, words such as "lux", "gratis", "withdraw", "euro", "gene" or expressions like "psychological disorder", "return to prehistory" are present in texts annotated with the labels grouped under *agency*.

For the **French** subset, we noticed similar patterns for the terms linked to instances containing stereotypes, with links to violence ("knife"), but also to school ("schooling", "student"), which are often brought up in instances labeled with discredit under the *agency* group (more particularly, DD), in claims that children of immigrants receive disproportionate financial aid from the state. For instances which do not contain stereotypes, we notably find terms related to misinformation ("fake", "fake news", "ridiculous"), which are often levelled against tweets containing stereotypes linked to racial hoaxes.

This underlines the polarization found in the reactions to RHs, by which one section of the users oppose ideas embodied in the RH since they are spread by a proven fake news, thereby avoiding playing the game of attributing certain characteristic to the population designated by the label of

[6]The n-grams are weighted using the TF-IDF measure on normalized texts; the phase of preprocessing involved: the deletion of all user mentions, stop-words, punctuation and URLs, leaving only words that were lexically significant; the tokenization, and the lemmatization with the SpaCy library.

"immigrants"; while another section of the users deliberately ignores the fact that the news has been diverted to focus on the designation of immigrants as the source of a problem. Immigrants are blamed either by their mere presence, which would represent a competition for limited resources, or by their acts, essentializing them as individuals all alike, violent and imposing their foreign culture.

The **Spanish** dataset also present interesting patterns in line with the topics of the grouped data. Firstly, the most prevalent words from texts with no stereotypes are related mainly to politics and economy ("unemployment", "reform", "communist"), whereas texts containing stereotypes show representative words used in RHs ("illegal immigrant", "tradition", "pay health care"). In relation to the categories of discredit, the main characteristic of *communion* is the perception of immigrants as violent, but also as victims, a fact that we can observe in the words like "invasion", "security", "serious" on the negative view, and "poor", "foreigner" and "right" on the patronizing view. On the other hand, *agency* takes a rather derogatory point of view of immigrants, which is displayed in words such as "idiot", "inferior race" and "dumb". The lexica in all languages reflect the stereotypes used against immigrants and the different forms of discredit.

## 6 Conclusion and Future Work

In this paper, we presented the first outcomes of a study of the stereotypes that are spread through racial hoaxes, with the aim of creating NLP resources and tools to automatically detect them. In order to address this challenging task, we started with an examination of the psychological and computational literature on fake news and stereotypes. This helped us to build the MRHC, the first multilingual corpus of racial hoaxes, which includes RHs in Italian, Spanish, and French, classified according to the topic of the news they spread. We designed a multi-layered annotation scheme for the annotation of racial stereotypes that takes into consideration the conversational thread extracted from social media. We applied it for the first time to a newly created multilingual dataset of Twitter reactions to RHs. Thanks to the outcomes of the annotation procedure, we were able to perform cross-cultural and cross-language analyses of these texts that are shaped in a Twitter conversational structure.

The results show that the presence of stereotype is, in general, lower within the RHs domain, with respect to its percentage in other pre-existing more general-purpose datasets (e.g., the ones developed within shared tasks). Other relevant findings show that, even if the first source RH contains a stereotype, in the following replies in the conversational thread, the presence of stereotypes decreases. Additionally, the dimension of implicitness was shown to be highly dependent on the dimension of contextuality in this domain. Content-wise, from an observation of RHs' topics, crossed with a lexical analysis (counting the most relevant tokens and expressions in each language subset), the outcomes show how the presence of stereotypes is linked to words that are typically grounded within the specificities of a certain language or culture. Finally, it can be observed that people who continue to spread a stereotypical view, originated in the source tweet and throughout the replies-to-replies, typically use polarized expressions that are in line with the original RH that generated the full conversational thread.

Thanks to the resources and framework elaborated in this study, it will be possible to investigate the spread of racial stereotypes on social media in a finer-grained way from a computational perspective and in a multilingual context. Furthermore, these steps are essential for developing computational tools for the automatic detection and classification of racial stereotypes in real-life scenarios.

## Limitations

In this work we presented, for the first time, a multi-layered scheme for the annotation of racial stereotypes in social media data in three different languages and in conversational threads. This work can, therefore, be considered pioneering and its multi-layered annotation scheme might require adaptation if applied to datasets with very different characteristics. The Stereotype Content Model inspired the annotation and analysis of stereotypes, by providing a socio-psychological theoretical framework. However, when being as faithful as possible to it during the annotation process, a computational setting can benefit from the integration of a more data-driven perspective.

Furthermore, the three subsets of the multilingual dataset of "reactions to racial hoaxes" now have very different sizes and present many unbalanced dimensions and high data sparsity. If in the

future they will be used for computational tasks, as it is intended, they should be made more balanced and more inclusive in terms of data sources.

Finally, cultural and geographical differences between the three languages of this study need to be taken into account and investigated in a deeper fashion, as it emerged that they are not trivial.

## Ethics Statement

The authors have carefully considered the ethics of conducting this kind of study regarding racial stereotypes on social media, and include here their assessment of the ethical issues raised and how to approach them. Throughout the project, they will be critically reflexive about unanticipated ethical issues arising from its sensitive, qualitative and digital nature.

The research presented in this work does not include any studies with human participants carried out by any of the authors. Furthermore, the data that was used is textual content from social media extracted from datasets publicly available to the research community and which also conform to the Twitter Developer Agreement and Policy, which allows for the unlimited distribution of the numeric identification number of each tweet. Each tweet in Italian, Spanish, and French has been translated and adapted into English in order to ensure the anonymity of the author.

—

Hiring policy: beside the authors of this article, other researchers were involved. We hired two Italian native-speaking annotators (one male and one female: a master's degree student in Linguistics, and a pre-doctoral student). We hired two Spanish native-speaking annotators (one male and one female, both Linguistics undergraduate students in their last year). We hired three French native-speaking annotators (three females, two master's degree students in "Linguistics, Communication and Gender", and one Linguistics undergraduate student). All hired workers have either received a monetary compensation or university credits valid for their career.

## Acknowledgements

This work is supported by the International 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 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. Furthermore, the authors would like to acknowledge Francesca D'Errico, Marinella Paciello, Giuseppe Corbelli, Paolo Giovanni Cicirelli and Concetta Papapicco, who contributed to the definition of the theoretical framework for the annotation of racial stereotypes and to the data collection procedure.

## References

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. *The nature of prejudice*. Addison-wesley Reading, MA.

Alejandro Ariza-Casabona, Wolfgang S. Schmeisser-Nieto, Montserrat Nofre, Mariona Taulé, Enrique Amigó, Berta Chulvi, and Paolo Rosso. 2022. Overview of DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish. *Procesamiento del Lenguaje Natural*, 69:217–228.

Camiel J Beukeboom and Christian Burgers. 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.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Andrea Cerase and Claudia Santoro. 2018. *From racial hoaxes to media hypes: Fake news' real consequences.*, pages 333–354. Amsterdam University Press.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 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*, pages 2833–2844, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amy J.C. Cuddy, Susan T. Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology*, 40:61–149.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

Mattias Ekman. 2019. Anti-immigration and racist discourse in social media. *European Journal of Communication*, 34(6):606–618.

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

Susan T. Fiske, Amy J.C. Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83.

Susan T. Fiske and Shelley E. Taylor. 2013. *Social cognition: From brains to culture*. Sage.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2022. Computational modeling of stereotype content in text. *Frontiers in artificial intelligence*, 5.

Reynier Ortega-Bueno, Berta Chulvi, Francisco Rangel, Paolo Rosso, and Elisabetta Fersini. 2022. Profiling Irony and stereotype spreaders on Twitter (IROSTEREO) at PAN 2022. *CEUR-WS. org*.

Concetta Papapicco, Isabella Lamanna, and Francesca D'Errico. 2022. Adolescents' Vulnerability to Fake News and to Racial Hoaxes: A Qualitative Analysis on Italian Sample. *Multimodal Technologies and Interaction*, 6(3):20.

Matthijs Rooduijn, Bart Bonikowski, and Jante Parlevliet. 2021. Populist and nativist attitudes: Does ingroup-outgroup thinking spill over across domains? *European Union Politics*, 22(2):248–265.

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

Juan Javier Sánchez-Junquera, Berta Chulvi, Paolo Rosso, and Simone Paolo Ponzetto. 2021a. How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. *Applied Sciences*, 11(8).

Juan Javier Sánchez-Junquera, Paolo Rosso, Manuel Montes, Berta Chulvi, et al. 2021b. Masking and BERT-based models for stereotype identication. *Procesamiento del Lenguaje Natural*, 67:83–94.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Antonio Stranisci, Cristina Bosco, Caselli Tommaso, Viviana Patti, Russo Irene, et al. 2020. HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–9. CEUR.

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

Claire Wardle and Hossein Derakhshan. 2018. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. *Ireton, Cherilyn; Posetti, Julie. Journalism, 'fake news' & disinformation. Paris: Unesco*, pages 43–54.