SteRHeotypes Project. Detecting and Countering Ethnic Stereotypes emerging from Italian, Spanish and French Racial hoaxes

Francesca D'Errico[1], Cristina Bosco[2], Marinella Paciello[3], Farah Benamara[4,5], Paolo Giovanni Cicirelli[1], Viviana Patti[2], Véronique Moriceau[4] and Mariona Taulé[6]

[1] *University of Bari 'Aldo Moro', Via Crisanzio 42, Bari, Italy*
[2] *University of Turin, Corso Svizzera 185, Torino, Italy*
[3] *Uninettuno Telematic International University, Corso Vittorio Emanuele II, Roma, Italy*
[4] *IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France ;* [5] *IIPAL, CNRS-NUS-ASTAR, Singapore*
[6] *Universitat de Barcelona - CLiC, Gran Via de Les Corts Catalanes 585, Barcelona, Spain*

Abstract

One of the Challenges for Europe should be to understand how media and fake news can reinforce prejudices against immigrants. The main objective of the STERHEOTYPES (STudying European Racial Hoaxes and sterEOTYPES) project, led by the 'Aldo Moro' University of Bari and funded by the Fondazione Compagnia di San Paolo ("Challenges for Europe"), is to promote awareness of the psychological processes generated by racial hoaxes, in the new digital generations, in three European Mediterranean countries: Italy, Spain and France. This paper presents the main findings aimed at (i) Understanding and countering factors associated to online racial hoaxes and online stereotypes through psycho-educational interventions in schools using an ad hoc designed web-app; and (ii) Investigating the features and dynamics of racist stereotypes through computational analysis based on novel annotated datasets that have been used to train automatic stereotype detection models in multilingual and multicultural perspectives.

Keywords

Racial misinformation, Stereotypes detection, Multilingual annotated corpora, Implicitness, Socio-Psychological models

1. Introduction

The international Project STERHEOTYPES (https://www.irit.fr/sterheotypes) starts from the consideration that one of the main Challenges to be addressed in European countries should be to understand and promote awareness on how media and fake news can reinforce prejudices against refugees and immigrants, especially in the Mediterranean area. In this context the levels of prejudice of 'common people' could increase when fed by 'fake' information, as in the case of so-called 'racial hoaxes' [1]. Racial Hoaxes (RHs) are communicative acts created to spread information regarding alleged threats to someone's health or safety by individuals or groups because of race, ethnicity or religion. In relation to RHs, the literature has considered people's short-term emotional reactions while neglecting socio-psychological processes [2], such as stereotypes and prejudices, which can reinforce anti-immigrants' attitudes [3] especially considering their impact on young 'digital natives'. In this sense, the project seeks to promote, first of all, civic awareness for future European citizens by contributing to the development of their social resilience to pervasive disinformation

SEPLN-CEDI2024: Seminar of the Spanish Society for Natural Language Processing at the 7th Spanish Conference on Informatics, June 19-20, 2024, A Coruña, Spain

paolo.cicirelli@uniba.it (P. G. Cicirelli); francesca.derrico@uniba.it (F. D'Errico); cristina.bosco@unito.it (C. Bosco); marinella.paciello@uninettunouniversity.net (M. Paciello); farah.benamara@irit.fr (F. Benamara); viviana.patti@unito.it (V. Patti); veronique.moriceau@irit.fr (V. Moriceau); mtaule@ub.edu (M. Taulé)

0000-0002-8957-665X (F. D'Errico); 0000-0002-8857-4484 (C. Bosco); 0000-0002-9023-0656 (M. Paciello); 0000-0002-0685-1864 (F. Benamara); 0000-0001-5002-4834 (P. G. Cicirelli) 0000-0001-5991-370X (V. Patti); 0000-0001-9641-0714 (V. Moriceau); 0000-0003-0089-940X (M. Taulé)

on immigration by fostering their prosocial attitude and sense of solidarity by means of a cross-cultural approach. Therefore, the main aim of the STERHEOTYPES project, is to investigate the stereotypes against immigrants, to promote awareness related to the psychological processes generated by racial hoaxes, particularly, in new digital generations, by means of integrated methodological approaches coming from psychology and computational linguistics, observing the multilingual and multicultural scenario in three European Mediterranean countries: Italy, Spain and France. In particular, the psychological teams, University of Bari (UniBari) and Uninettuno University (Utiu), performed experimental studies in real educational contexts with the aim of understanding the socio-psychological processes arising from online RHs by considering individual differences related to prejudice and their potential effects on emotional and self-regulative processes within online communicative contexts. The three computational linguistics teams involved in the project, Università di Torino (UniTo), Universitat de Barcelona (UniBa) and Université du Toulouse (UniTou), worked instead on the development of linguistic resources and tools for the automatic detection of stereotypes. In particular, the three groups defined a common methodology for the collection and a shared scheme for the fine-grained annotation of the collected data, inspired by psychological studies (mainly Stereotype Content Model [4]). Using this methodology and applying the schema, they created a series of comparable corpora that enabled computational experiments in a multilingual environment, but also linguistic analyses. The main research questions the STERHEOTYPES project would like to address are the following:

- RQ1: What are the psychological features (values, self-efficacy, analytical reasoning, prejudice) that make young people more easily influenced by racial hoaxes?
- RQ2: Can an ad hoc intervention lead to adolescents becoming aware of media biases and proactive in promoting intergroup contact?
- RQ3: What are the linguistic patterns that emerged from the expressions of ethnic stereotypes in the different languages studied?
- RQ4: Is it possible to automatically detect stereotypes?

2. Scientific and Technical results

2.1. Theoretical framework and definition of cases and tools

The project (led by UniBari) started with literature review on RHs aimed at selecting and classifying several types of RHs according to their focus on the main different actors involved in hosting immigrants that can negatively influence attitudes. UniBari contributed to consolidating the project's psycholinguistic framework in relation to the potential linguistic features to take into account, such as cognitive stereotypes, prejudices, stance, emotional responses and attitudes toward immigrants (RQ1). This led UniTo and UniBari to release a first guidelines for the annotation of stereotypes to be applied to multilingual corpora extracted from social media. Meanwhile, Utiu focused on the definition of psychological variables and evaluation tools to be used for the experimental part of the project, in order to understand the role of individual differences in moderating media effects on adolescents' responses and in maintaining or in changing anti-immigrants' attitudes/stances (RQ2).

2.2. Data and corpus collection and Analysis

To analyze and understand the stereotypes, prejudices and emotional reactions generated by RHs in a cross-cultural perspective (RQ3), and to identify the way in which these stereotypes are linguistically expressed and to be able to detect them automatically (RQ4), the following corpora were created:

- The RacialHoax corpus consists of 239 RHs (70 in French, 97 in Italian and 72 in Spanish) related to immigration which were manually extracted from French, Italian and Spanish fact-checking websites or newspaper articles verifying or refuting claims made in social media. Then, we used the items in these corpora as seeds for collecting conversations, which were reactions to these hoaxes, retrieved from Twitter to create the StereoHoax corpus [5]. Regarding the annotation of the RacialHoax corpus, we manually classified each RH into six categories in accordance with the main topic they addressed ('Benefits', 'Security', 'Migration control', 'Culture/Religion', 'Public Health' and 'Others'). We also annotated other categories such as the Target of the RH, Stance, Implicitness (i.e. whether the stereotype was expressed implicitly or explicitly), the event described, and for each of these categories the words associated to them that are crucial for classifying these categories. D'Errico and colleagues [1] considered these initial categories and, in addition, the annotation of further psycholinguistic features (such as whether the immigrants are described as the subject or object of the RH, the verbal form, types of adjectives and the presence of affective words) to conduct a deeper psycholinguistic analysis of the Italian RHs.

The StereoHoax corpus consists of 17,814 tweets (9,342 in French, 3,123 in Italian and 5,349 in Spanish) reporting on and responding to racial or ethnic hoaxes about immigrants dated from 2019 to 2022. This corpus was also manually annotated with the presence or absence of 'Stereotype', 'Contextuality' (to indicate whether the conversational context is needed to interpret the stereotype), 'Implicitness' (explicit or implicit stereotype) and the forms of 'Discredit' described in Bosco and colleagues [6] (benevolence, dominance up, affective competence, dominance down, competence and physical based on the Fiske's Stereotype Content Model [4, 7], The Spanish part of StereoHoax was also annotated with the 'Types of implicitness' following the proposal described in Schmeisser and colleagues [8] and the 'Stereotype category' applied in the DETESTS dataset [9]. The

Italian subset was also annotated with the 'Stance' expressed in the messages of the conversations towards the veracity of the hoax [10] adopting the SDQC schema [11] for indicating whether the author of the message 'Support', 'Deny', 'Query' or 'Comment' the veracity of the hoax, and adding the label 'Head' to identify the texts that spread the hoax or start the conversational thread. The Italian part of StereoHoax is also being annotated with 'Type of implicitness'. Finally, for the French part, in addition to annotating for the presence/absence of stereotypes and stance as well as types of discredit, we also annotated for the presence/absence of hate speech. This will be particularly useful for multitask learning experiments, where hate speech can help for stereotype detection and vice-versa (see Chiril and colleagues [12]).

The aim of these corpora is twofold: 1) to analyse the way in which stereotypes are expressed in the languages studied and 2) to be used as datasets for training and evaluating systems based on machine learning models for detecting and classifying stereotypes. Several experiments were performed using these corpora. For instance, in Bourgeade and colleagues [5], a first cross-lingual comparative analysis of StereoHoax was presented, focusing on the interaction between the topics of RHs, stereotypes and the forms of discredit expressed in the messages of StereoHoax in French, Italian and Spanish. Cignarella and colleagues [10] analyse the distribution of stance labels and their interrelationship with stereotypes providing information about the way in which the structure and nature of conversations and lexical choices in messages may affect the perceived stance of the user to RHs. Schmeisser-Nieto and colleagues [13] describes BERT-based models designed to detect stereotypes related to immigrants trained with the Spanish subset of the StereoHoax corpus. Predictions from GPT-4 are also generated and analised. The aim is to provide insight into linguistic distinctions in the way humans and these models perceive stereotypes. The results obtained show the need to refer to the conversational context to interpret the stereotype and the difficulty to identify implicit stereotypes, particularly, for models. The StereoHoax corpus in Spanish will be used in the DETESTS-Dis task at IberLEF 2024 (https://detests-dis.github.io/).

2.3. Experimental psychological studies in natural settings

The psychological units collected a sample of real data produced by adolescents by contacting Italian schools and performing a data analysis aimed at the definition of individual psychological dimensions (personal values, empathy levels, anti-immigrant attitude and implicit prejudice) that could moderate the responses of young people to RHs. We also collected spontaneous comments by the adolescents reflecting their reaction to stimuli on emotions and attitudes to RHs about immigrants. In a first step, UniBari and Utiu collected data by involving more than 2600 students mainly adolescents by means of focus groups to test their knowledge about RHs and fake news. Then, they tested measures and RHs that can affect digital natives' stereotypes and also their emotions in Italian contexts (Rome, Bari). Finally, they engaged with a web-app called 'ROLLING MINDS!' [14] which is aimed at testing adolescents' comments and reactions to RHs and misleading news. In cooperation with the computational linguists at UniTo, the psychological units developed an intervention procedure focused on so-called media biases [15]. More than 600 Italian students were involved, first in a simulated version and then by means of a conversational approach to recognize linguistic stereotypes and anti-immigrant narratives. The conversational web-app ROLLING MINDS!' will be translated in Spanish and English in collaboration with UniBa to collect longitudinal and cross-cultural data. The Italian data collected across several studies was also analysed in relation to students' data to propose a first definition of the socio-psychological processes underlying influence of RHs [16, 17, 18, 19].

2.4. Action research for promotion of awareness on racial hoaxes'

Both experimental and media findings were presented by UniBari and Utiu in schools to promote 'digital natives' awareness of the toxicity of RH [14, 20]. In particular, young people were invited to reflect on the emotional and cognitive processes involved in the discussion of RHs considering the source of messages, their media biases and immigrants' points of view [15]. Teachers and students were involved in several debriefings in which the researcher explained the theoretical framework guiding research and the psychological differences that could interact with mediated contexts. The researchers and other scholars in the field of misinformation, hate speech and online racism will also provide a video-course and comprehensive guidelines to the individuals involved and future teachers for promoting the diffusion of a preventive intervention by using the ROLLING-MINDS! web-app (Figures 1, 2) to promote self-reflection on personal aspects involved in the discussion of RHs. This video-course will be available on the project website at the end of the project. Finally, the web-app, thanks to an agreement with BDS design (which supported psychological units in the implementation of the web-app), will be used to monitor the trends of the psychosocial variables related to RHs throughout and after the end of the STERHEOTYPES project.

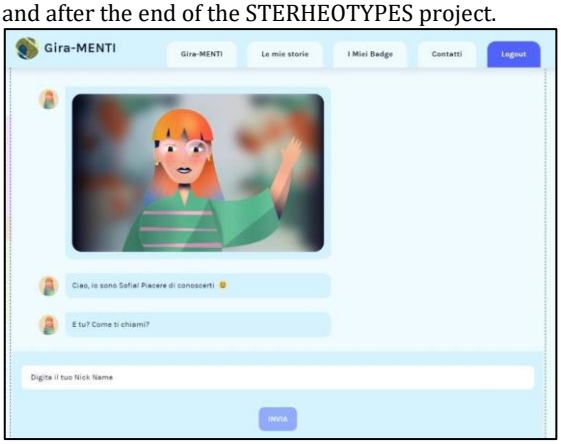

Figure 1: ROLLING-MINDS! web-app.

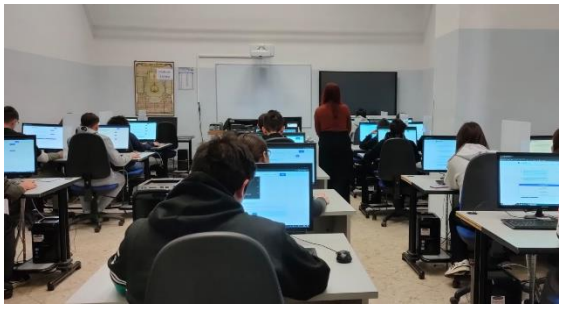

Figure 2: ROLLING-MINDS! at school.

2.5. Science communication measures

All the meetings and scientific communications across reciprocal domains were fully documented at the project website, where both researchers and schools involved in the experimental part and other possible stakeholders (teachers, journalists, policymakers) can deepen these topics. https://www.irit.fr/sterheotypes/publications. As to the performed synergies, the project team was composed by senior scholars, all females, skilled in Psychology and Computational Linguistics. They involved, thanks to the project, 10 among pre and post-doc junior researchers under their supervision; They, at different stages of their academic career were supported by the Project, in particular dr Cignarella and Frenda from UniTorino, dr. Corbelli from UniNettuno, Sebastian Wolfgang Schmeisser UniBarcelona, Paolo Giovanni Cicirelli and Carmela Sportelli from UniBari, Tom Bourgade from UniToulose will discuss their PhD Thesis on the project theme, by disseminating their results across academic communities coming from Psychology and Computational linguistics.

3. Conclusions

Regarding the psychological RQ1, we showed across studies that promoting an analytical and self-regulated approach to the reading of RHs led adolescents to be more aware and prone to intergroup contact. These considerations are based on results emerging from several sessions with the conversational web-app 'ROLLING MINDS!' [14, 20]. The remaining months at the end of the project will take advantage of the experience gained in these three years to, in the case of the psychological units, collect data and reactions to RHs by adolescents, in order to allow the working group of computational linguists to carry out a detection of stereotypes in a natural context, making a comparison between detection in social media and in real contexts, by also taking into account socio-psychological variables.

This effort allows us to create a new dataset named STERHEOSCHOOL (currently under review) that will be presented in future European congresses on these themes. Finally, the implementation of the web-app has demonstrated good results in terms of prejudice prevention [14, 20], and has undergone some variations which will be further tested in Italian schools. Furthermore, the translation by the UniBa will allow the web-app to be applied in other contexts (Spanish and English in particular). The results of the subsequent surveys will allow us to further understand the risk and protective factors of racial misinformation in adolescence. The outcomes of the project allow us to pave the way for further investigations in the area of stereotype detection. Regarding resources (RQ3), we are working on the development of a novel corpus annotated with the same scheme applied to the StereoHoax corpus, but including texts collected in the experiments organized by the team of psychologists in schools, which represent a genre that is different but characterized by an important similarity, particularly the collection with a conversational context. The availability of comparable multilingual corpora is currently enabling the application of data augmentation techniques based on an approach that includes machine translation and back translation. This will help us to address the limited size of the currently existing data sets for the three languages of the project. Regarding the development of tools for stereotype detection (RQ4), we are continuing our research on modeling this complex phenomenon whose detection can be especially challenging.

Acknowledgements

This work was supported by the European project 'STERHEOTYPES—STudyingEuropean Racial Hoaxes and sterEOTYPES' funded by 'Challenge for Europe' call for Project, Fondazione Compagnia di San Paolo and the Volkswagen Stiftung (CUP: B99C20000640007). We extend our gratitude to the BSD Design team for their technical assistance and support in the development of the web app "Rolling Minds".

References

[1] F. D'Errico, C. Papapicco, M. Taulé, 'Immigrants, Hell on Board'. Stereotypes and Prejudice Emerging from Racial Hoaxes through a Psycho-Linguistic Analysis, Journal of Language and Discrimination, 6(2) (2022): 191–212.

[2] F. D'Errico, M. Paciello, Online moral disengagement and hostile emotions in discussions on hosting immigrants, Internet Research, 28(5) (2018):1313–1335.

[3] C. Wright, R. Brinklow-Vaughn, K. Johannes, F. Rodriguez, (2021). Media Portrayals of Immigration and Refugees in Hard and Fake News and Their Impact on Consumer Attitudes, Howard Journal of Communications, 32(4) (2021): 331–351.

[4] S. T. Fiske, A. J. C. Cuddy, P. Glick, J. Xu, A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology, 82(6), (2002):878–902.

[5] T. Bourgeade, A.T. Cignarella, S. Frenda, M. Laurent, W.S. Schmeisser-Nieto, F. Benamara, C. Bosco, V. Moriceau, V. Patti, M. Taulé, A

Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads, in: Findings of the Association for Computational Linguistics, EACL 2023, 2023, Croatia, pp.674-684. doi: 10.18653/v1/2023.findings-eacl.51

[6] C. Bosco, V. Patti, S. Frenda, A.T. Cignarella, M. Paciello, F. D'Errico, Detecting Racial Stereotypes: An Italian Social Media Corpus where Psychology Meets NLP, Information Processing and Management, 60(1) (2023): 103118.

[7] F. D'Errico, I. Poggi, L. Vincze, Discrediting signals. A model of social evaluation to study discrediting moves in political debates, Journal on Multimodal User Interfaces, 6 (2012):163-178.

[8] W.S. Schmeisser-Nieto, M.Nofre, M. Taulé, Criteria for the Annotation of Implicit Stereotypes, in: Proceedings of the 13th Conference on Language Resources and Evaluation, LREC 2022, Marseille, 2022, pp. 753–762.

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

[10] A.T. Cignarella, S. Frenda, T. Bourgeade, C. Bosco, F. D'Errico, Linking Stance and Stereotypes About Migrants in Italian Fake News, in: Proceedings of CLiC-it 2023 19th Italian Conference on Computational Linguistics, CEUR Ws-org Vol-3596, Venice, Italy, 2023.

[11] A. Aker, L. Derczynski, K. Bontcheva, Simple Open Stance Classification for Rumour Analysis, in: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017), INCOMA Ltd., 2017.

[12] P. Chiril, F. Benamara, V. Moriceau, "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification?. In: Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 2833-2844).

[13] W.S Schmeisser-Nieto, Human vs. Machine Perceptions on Immigration Stereotypes, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024, Torino, 2024.

[14] F. D'Errico, P.G. Cicirelli, G. Gorbelli, M. Paciello (2024) Rolling Minds. a Conversational Media to Promote Intergroup Contact by Countering Racial Misinformation Through Socio-Analytic Processing in Adolescence.

[15] R. Paul, L. Elder, The thinkers guide for conscientious citizens on how to detect media bias and propaganda in national and world news: Based on critical thinking concepts and tools. Rowman & Littlefield, 2004.

[16] F. D'Errico, G. Corbelli, C. Papapicco, M. Paciello, How Personal Values Count in Misleading News Sharing with Moral Content, Behavioral Sciences, 12(9) (2022), 302. doi: 10.3390/bs12090302

[17] G. Corbelli, P.G. Cicirelli, F. D'Errico, M. Paciello, Preventing Prejudice Emerging from Misleading News among Adolescents: The Role of Implicit Activation and Regulatory Self-Efficacy in Dealing with Online Misinformation, Social Sciences,12(9) (2023):470.

[18] C. Papapicco, I. Lamanna, F. D'Errico, 'Adolescents' vulnerability to False Information and to Racial Hoaxes. A qualitative content analysis on Italian sample, Multimodal Technologies and Interaction, 6(3) (2022): 20. doi:10.3390/mti6030020

[19] M. Paciello, G. Corbelli, F. D'Errico, The Role of Self-efficacy Beliefs in Dealing with Misinformation among Adolescents, Frontiers in Media Psychology, vol. 14, (2023).

[20] F.D'Errico, P.G. Cicirelli, G. Corbelli, M. Paciello, Addressing racial misinformation at school: a psycho-social intervention aimed at reducing ethnic moral disengagement in adolescents, Social Psychology of Education, (2023). doi:10.1007/s11218-023-09777-z