Information Processing and Management 60 (2023) 103118

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP

Cristina Bosco [a,*], Viviana Patti [a], Simona Frenda [a], Alessandra Teresa Cignarella [a], Marinella Paciello [c], Francesca D'Errico [b]

[a] Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, 10149 Torino, Italy
[b] Dipartimento Formazione, Psicologia, Comunicazione, Università di Bari "Aldo Moro", Via Scipione Crisanzio 42, 70122 Bari, Italy
[c] Facoltà di Psicologia, Università Telematica Internazionale UniNettuno, Corso Vittorio Emanuele II, 39, 00186, Rome, Italy

ARTICLE INFO

Keywords:
Stereotypes
Social psychology
Natural language processing
Social media
Lexical analysis
Corpus
Linguistic annotation
Machine learning
BERT

ABSTRACT

The generation of stereotypes allows us to simplify the cognitive complexity we have to deal with in everyday life. Stereotypes are extensively used to describe people who belong to a different ethnic group, particularly in racial hoaxes and hateful content against immigrants. This paper addresses the study of stereotypes from a novel perspective that involves psychology and computational linguistics both. On the one hand, it describes an Italian social media corpus built within a social psychology study, where stereotypes and related forms of discredit were made explicit through annotation. On the other hand, it provides some lexical analysis, to bring out the linguistic features of the messages collected in the corpus, and experiments for validating this annotation scheme and its automatic application to other corpora in the future. The main expected outcome is to shed some light on the usefulness of this scheme for training tools that automatically detect and label stereotypes in Italian.

1. Introduction

In the last few years, media have often raised local and international social issues and events concerning immigration, such as the war in Syria, the persistent state of crisis in many African countries, and the subsequent impulse to migration fluxes in Europe. Newspapers, TV channels and social networks offer news, information and opinions related to multi-ethnic relationships. They may also spread a racist and discriminatory discourse based on ethnic prejudices and stereotypes, for instance, hoaxes focused on racial differences. Sometimes such forms of communication also include Hate Speech (henceforth *HS*) and compel us to think about and deal with issues related to social conflicts.

A variety of answers to these societal challenges is proposed by policymakers, governments and civil society organizations, but also by researchers involved in local and European projects for monitoring, preventing and countering racism and xenophobia spreading in our society.[1]

For contrasting HS, the effort of researchers from several disciplinary fields is currently focused on mass media and on different facets of the language used in these contexts. Data network analysis, social psychology and discourse analysis are devoted to studying

* Corresponding author.
E-mail addresses: cristina.bosco@unito.it (C. Bosco), viviana.patti@unito.it (V. Patti), simona.frenda@unito.it (S. Frenda), alessandrateresa.cignarella@unito.it (A.T. Cignarella), marinella.paciello@uninettunouniversity.net (M. Paciello), francesca.derrico@uniba.it (F. D'Errico).

[1] See, for instance, *Hatebase* http://hatebase.org/, *Mandola* http://mandola-project.eu/, *Coalition of Positive Messengers to Counter Online Hate Speech* co-funded by Rights, Equality and Citizenship/Justice Programme of the European Union, *Building Respect on the Internet by Combating hate Speech BRICKS* https://www.bricks-project.eu/about-the-project/ funded by Fundamental Rights and Citizenship Programme of the European Union.

https://doi.org/10.1016/j.ipm.2022.103118
Received 15 July 2022; Received in revised form 15 September 2022; Accepted 10 October 2022
Available online 1 November 2022

and monitoring them; but the social issues they raise impose challenges and expectations especially for Artificial Intelligence in general and for Computational Linguistics in particular. Consequently in the last few years the detection of hateful contents in social media has been among the hottest topics for Natural Language Processing (NLP), text classification and opinion mining (Basile et al., 2019; Bosco, Dell'Orletta, Poletto, Sanguinetti, & Tesconi, 2018; Pang & Lee, 2008).

On the one hand, online communication, and in particular the so-called *user-generated contents* (UGC), offers us the largest amount of data ever seen before (i.e. big data) where HS and related phenomena are plentifully represented. On the other hand, several studies confirm that stereotypes, which are the cognitive basis of HS, are learned by various forms of socialization and especially by public discourse, spoken and written within mass media, and interpersonal conversations influenced by such public discourse (van Dijk, 2016), like those collected in social networks. The notion of HS is related in literature with that of discrimination and stereotype (Bauwelinck & Lefever, 2019; van Dijk, 2016; Fiske, 1998), which are the cognitive and behavioral counterparts of this phenomenon in human social life.

The first objective of this paper is to introduce a novel corpus of Italian social media texts built within the context of social psychology research (D'Errico & Paciello, 2018) aimed at exploring the socio-cognitive mechanisms underlying the opposition to immigrants' hosting, and then annotated for making explicit stereotypes against immigrants and the related discredit forms. Nevertheless, beyond the opportunity that this project offers to deepen the theme of multi-ethnic relationships according to social psychology, it also allowed us to pursue a second objective which is to provide some contribution to the advancement of computational linguistic research about the possibility of automatically detecting and annotating stereotypes in texts.

This corpus, exploited – in a previous release – within social psychology studies,

is a collection of Facebook messages now made adequate also for training and testing NLP tools for Italian. This effort is coordinated by the Department of Formation, Psychology and Communication of the University of Bari "Aldo Moro" and also forms part of the Hate Speech Monitoring program of the Computer Science Department of the University of Turin[2] with the aim at detecting, analyzing and countering HS implementing an inter-disciplinary approach (Bosco et al., 2017) within the context of the international project STERHEOTYPES[3].

We can summarize the main steps of the methodology applied in this study as follows. First of all, we discuss the design and application of a novel fine-grained annotation scheme. This step allows us to thoroughly analyze how, and using which specific discredit forms, people express racial stereotypes by referring to a real social context and specific situation where discrimination and racism arise. The main goal of this paper is indeed to pave the way for the improvement of linguistic resources and tools for automatic stereotype detection and annotation, but also to show how to take into account other aspects involved in the generation of hateful content in order to give evidence to multiple facets of the hateful communication.

Secondly, for validating the scheme after its application on the FB-Stereotypes corpus,[4] we perform a lexical analysis where word n-grams collected from the dataset are observed and compared with those drawn from other smaller datasets annotated with the same scheme. In particular, we exploited a sample of tweets collected as reactions to a set of racial hoaxes and another sample from a benchmark corpus of an HS detection shared task.

Finally, we provide some experiments especially focused on the stereotype category: for evaluating the possibility of automatically labeling stereotypes in a novel set of data an automatic stereotype detection tool is indeed trained on FB-Stereotypes and these other cited resources.

The paper is organized as follows. The next section surveys related literature regarding stereotypes and the major computational experiences about the detection of stereotypes, HS and other alike phenomena in social media texts. The third section is focused on the collection of data included in the FB-Stereotypes corpus, while the fourth presents the annotation scheme we designed and applied for making HS, stereotype and related phenomena explicit. Finally, in Sections 5 and 6, a lexical analysis and computational experiments are presented and discussed as the results we achieved in this project.

2. Related work and background

The notions of stereotype and prejudice are often used almost as synonym terms and there exists a close relationship between them that motivates this common use. The stereotype is indeed the cognitive nucleus of prejudice, which assumes in turn the face of discrimination, racist and hateful behaviors in social interactions, such as HS.

A **stereotype** consists of a firmly held association between a social group and some features, like physical, mental, behavioral or occupational quality, e.g. "*blondes are ditzy bimbos*", "*engineers are geeks*". It mostly consists of a generalization about a group of people, in which selfsame characteristics are assigned to virtually all members of the group, regardless of the actual variation among the members (Allport, 1954; Aronson, Wilson, & Akert, 2013). On the one hand, the generation of stereotypes is the outcome of a very commonly applied and automatic mental process, that is categorization. On the other hand, stereotypes are acquired by humans during socialization or very often employing mass media (D'Errico, Papapicco, & Taulë Delor, 2022; Vaes, Latrofa, Suitner, & Arcuri, 2019)

[2] http://hatespeech.di.unito.it

[3] International project *STERHEOTYPES-Studying European Racial Hoaxes and Stereotypes* funded by VolksWagen Stiftung/Compagnia di San Paolo for the call for projects 'Challenges for Europe', https://www.irit.fr/sterheotypes/.

[4] A corpus of Italian posts containing stereotypes, extracted from Facebook. The description of the corpus collection and annotation is provided in Sections 3 and 4.

Both positive and negative stereotypes can be the basis for the development of a (positive vs. negative) **prejudice** about a social group, which consists of a specific behavioral attitude against a group or some member of it. Prejudice is then an evaluative attitude, composed of both cognitive and emotional factors, that in turn can be considered the base of verbal forms of racism or discriminative behavioral expressions, e.g. discrimination (Brown, 2011; van Dijk, 2016) which is expressed in various linguistic forms of **discredit**.

The majority of the psychosocial studies on immigrant stereotypes are focused on everyday and real-life interactional contexts (Durrheim, 2012), experimental contexts or classical media like journals (Vaes et al., 2019), while the contexts of social media, which are more and more pervasive in the construction of our attitude (Fields, 2016), have been strangely neglected until now. In this perspective, integrating psychological models of stereotypes within the linguistic automatic detection seems increasingly crucial to improve results in the detection of HS. Until now, only one paper proposes a taxonomy and a dataset of stereotypes for Spanish in a computational perspective (StereoImmigrants) (Sánchez-Junquera, Chulvi, Rosso, & Ponzetto, 2021); while only one dataset which includes the annotation of the presence vs. absence of stereotypes against immigrants has been made available, i.e. that used within the context of the Evalita evaluation campaigns for Italian NLP tools and resources as a benchmark for the shared tasks regarding *Hate Speech Detection* (HaSpeeDe).[5]

HS can be broadly defined as any expression "*that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people based on their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth.*" (Erjavec & Kovačič, 2012). Although definitions and approaches to HS vary considerably and depend on the juridical tradition of the country, many agree that what is identified as such cannot fall into the protection granted by the right to freedom of expression, and must be prohibited.

Automatic online HS identification requires multidisciplinary approaches and knowledge from different fields (like social psychology, law and social sciences), but the role that NLP plays seems to be crucial. The development of high-accuracy automatic tools able to identify HS is indeed assuming the utmost relevance due to the wide variety of its practical applications. Moreover, several events and shared tasks took place in the recent past or are currently on going, thus reflecting the interest in HS and HS-related topics by the NLP community. For instance, HS detection has been one of the most participated tasks of the international evaluation campaign *SemEval 2019*: the *Shared Task 5 on Hate Speech Detection against Immigrants and Women* for English and Spanish (Basile et al., 2019).[6] As far as Italian, a task about HS has been proposed also in *Evalita 2018*, i.e. *Hate Speech Detection* (HaSpeeDe) held in 2018 (Bosco et al., 2018), and cited above. Other related events, just to name a few, are the first and second edition of the *Workshop on Abusive Language*[7] (Waseem, Chung, Hovy, & Tetreault, 2017); the *First Workshop on Trolling, Aggression and Cyberbullying* (Kumar, Ojha, Zampieri, & Malmasi, 2018), that also included a shared task on aggression identification; the tracks on *Automatic Misogyny Identification* (AMI) (Fersini, Rosso, & Anzovino, 2018) and on *Authorship and aggressiveness analysis* (MEX-A3T) (Álvarez et al., 2018) proposed at the 2018 edition of *IberEval*, the *Automatic Misogyny Identification* task at *Evalita 2018* (Fersini, Nozza, & Rosso, 2018), and the task 6 of SemEval 2019 on *Identifying and Categorizing Offensive Language in Social Media* (OffensEval)[8] (Zampieri et al., 2019). For German the *GermEval Shared Task on the Identification of Offensive Language* (Wiegand, Siegel, & Ruppenhofer, 2018) has been recently followed by the *GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments* (Risch, Stoll, Wilms, & Wiegand, 2021) where the relationship among those different but closely related phenomena is also observed.

The high participation recorded by most of these shared tasks indicates the interest of the international community towards HS and encouraged the proposal of new editions of such contests, like (OffensEval 2: Multilingual Offensive Language Identification in Social Media (Zampieri, Nakov, Rosenthal, Atanasova, Karadzhov, Mubarak, Derczynski, Pitenis, & Çöltekin, 2020)),[9] the TRAC shared task (Kumar, Ojha, Lahiri, Zampieri, Malmasi, Murdock, & Kadar, 2020)[10] and the second edition of AMI[11] and HaSpeeDe, cited above.

All these tasks consist in discovering some kind of hateful content. Being these techniques mainly based on machine learning, they require the development of corpora which in most of cases are extracted from social media also considering a certain variety of sources. For instance, for the *Evalita 2018 HaSpeeDe* task, a couple of datasets has been arranged where texts respectively extracted from Facebook and Twitter have been provided together with their annotation making the presence of HS explicit. Several targets have been considered for building the Facebook dataset, which is extracted from another existing resource described in turn in Del Vigna, Cimino, Dell'Orletta, Petrocchi, and Tesconi (2017). The dataset from Twitter, described in Poletto, Stranisci, Sanguinetti, Patti, and Bosco (2017) and Sanguinetti, Poletto, Bosco, Patti, and Stranisci (2018), is instead focused on immigrants.

While several research activities are however reported about automatic HS detection, to our knowledge very few studies are about the automatic detection of racial stereotypes. Also in shared tasks providing datasets where they were annotated, systems were not properly tested for their ability in detecting this category, with the only exception of the HaSpeeDe shared task organized in 2020 (Sanguinetti et al., 2020) where a pilot subtask was about the detection of stereotypes. Only very recently some work has been issued about stereotype where a computational view is provided (Fraser, Kiritchenko, & Nejadgholi, 2022) and a task

[5] HaSpeeDe 2018: http://www.di.unito.it/~tutreeb/haspeede-evalita18/ and HaSpeeDe 2020: http://www.di.unito.it/~tutreeb/haspeede-evalita20 (Bosco et al., 2018; Sanguinetti et al., 2020).

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

[7] https://sites.google.com/view/alw2018/

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

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

[10] https://sites.google.com/view/trac2/shared-task

[11] AMI 2020: https://amievalita2020.github.io/.

Table 1
Distribution of *proself* and *prosocial* attitude towards immigrants in the collection of data done by the Bari's group referring to the Gianni Morandi's Facebook post.

Set	#comments
proself	7,046
prosocial	5,049
total	12,095

related to the detection of stereotype has been devised within PAN: *Profiling Irony and Stereotype Spreaders on Twitter* (IROSTEREO 2022).[12] An interesting analysis of the topic is also provided in Schmeisser-Nieto, Nofre, and Taulé (2022), where the implicitness of stereotype is especially observed. Moreover, while these cited works are in some broad sense about racial stereotypes (i.e. they are focused on stereotype against immigrants), some study addresses the particular typology of stereotypes which is related to gender and sexism (Chiril, Benamara, & Moriceau, 2021; Cryan et al., 2020) in line with the AMI shared tasks cited above.

We hypothesize that a deeper knowledge of phenomena strictly related to HS can be of some help in improving the quality of computational linguistics resources and the efficacy of automatic detection tools. This motivated the development of a novel and especially fine-grained annotation scheme centered on stereotype and its application on the case study, i.e. the **FB-Stereotypes**, and on some sample from other corpora as described in the following sections to perform some linguistic analysis and experiment.

This experience is in line with other annotation exercises in the field, where an annotation scheme is provided which was specifically designed to account for the multiplicity of facets that can contribute to the definition of HS (Sanguinetti et al., 2018), and to offer a broader tagset capable of better representing all those facets, which may contribute to increase, or rather mitigate, the impact of the message. This resulted in a scheme that includes, besides HS, aggressiveness, offensiveness, irony and stereotype.

Let us also highlight that this perspective is in tune with the one adopted in the new edition of *HaSpeeDe 2*, the Hate Speech Detection shared task for Italian proposed for EVALITA 2020, where the organizers chose to go beyond the simple binary classification (hateful vs. not-hateful), giving space also to finer-grained aspects pertaining, albeit indirectly, to HS, namely the presence of stereotypes referring to one of the targets identified within the task dataset (Muslims, Roma and immigrants). An error analysis of the best-performing systems participating in the HaSpeeDe 2018 dataset itself (Francesconi, Bosco, Poletto, & Sanguinetti, 2019) showed that the occurrence of these components constitutes a common source of error in HS identification.

3. Collecting the FB-Stereotypes corpus

The main reference corpus of this study, i.e. **FB-Stereotypes**, is a collection of 2,990 Italian messages retrieved from Facebook. It is the result of a filtering and selection applied on a larger pre-existing dataset (including 12,583 posts) collected thanks to a conjoint effort of the group of the University of Bari and that of the UniNettuno, for social psychology studies about *Emotions and Online Unethical dynamics toward immigrants hosting* (D'Errico & Paciello, 2018; D'Errico, Paciello, & Amadei, 2018).

This larger dataset was retrieved by referring to a social media case concerning a message posted by Gianni Morandi, a popular Italian singer, on April 21st, 2015. The post was about the victims of a severe shipwreck, that happened a few days earlier off the Sicily coast, that involved a boat with more than 700 migrants on board.[13] The singer aimed to encourage empathy towards those victims by equating migrants like them to the thousands of Italians who emigrated to America during the 20th century.

As it can be seen in Fig. 1, the message was composed of some text in Italian together with two images,[14] respectively representing a boat carrying Italians immigrating to the USA in the 20th century, and the boat involved in the shipwreck happened a few days earlier off the Sicily coast.

The University of Bari's group collected through Facebook API a total of 12,583 second-level comments were posted between April 21st to April 27th 2015 in response to Gianni Morandi's post. Comments addressed to Gianni Morandi, both positive and negative, the simple expression of agreement or disagreement and more than 400 comments containing links to videos or images were manually filtered. In particular, data pre-processing led to the identification of two main categories of messages: *prosocial* and *proself*, which were respectively in favor of or against hosting immigrants. They were subsequently annotated by two judges reaching a very good agreement (Cohen's $\kappa = 0.86$) and resulting in the distribution of proself and prosocial attitudes shown in Table 1.

FB-Stereotypes[15] is the novel corpus we have built on the top of this larger collection. By including only a portion of its data in the novel resource, we have built a smaller dataset which is however more suitable for a manual fine-grained annotation. By considering an equal proportion of data from the two categories annotated in the original corpus (prosocial and proself), we obtained a balanced corpus suitable for training and testing tools for the automatic detection of stereotype: 2,990 messages, that is 1,490 proself and 1,500 prosocial messages.

On the one hand, the main novelty featuring the FB-Stereotypes corpus, concerning the collection of data above depicted, is the fine-grained annotation especially designed to account for the stereotype and the relationships occurring between HS and stereotype.

[12] https://pan.bis.de/clef22/pan22-web/author-profiling.html#task-committee
[13] https://www.theguardian.com/world/2015/apr/19/700-migrants-feared-dead-
[14] https://www.facebook.com/giannimorandiofficial/posts/10153914629003438:0
[15] The corpus is available at https://github.com/boscoc/stereotypes.

Fig. 1. Original post from Gianni Morandi's Facebook wall.

On the other hand, what makes the dataset suitable for training and testing tools for the detection of stereotypes is the inclusion in the corpus of a comparable amount of messages marked as *proself*, i.e. expressing attitudes against letting refugees in Italy and hosting them, and marked as *prosocial*, i.e. those expressing attitudes favorable to hosting refugees in Italy. In Table 2 are displayed two examples per category of stereotype.

We expected to find a very high frequency of hateful expressions and negative stereotypes against immigrants inside the *proself* data sample, which could have been particularly useful for studying HS and related phenomena, while an almost total absence of them in the *prosocial* messages. Therefore we annotated all the 1,490 proself posts exploiting the annotation schema described in the next Section. Applying instead the annotation to a sample of 200 prosocial messages we validated the hypothesis that they do not include stereotype and we estimated it was not necessary to apply the annotation to the other prosocial ones. However, we consider this part of the corpus as a sort of "silver standard" where the proself part is a gold standard, as it was completely annotated and revised by humans, while the prosocial part was not.

Finally, this procedure allowed us to create a balanced corpus which includes positive and negative examples of the phenomenon observed, namely the racial stereotype and related forms of HS.

4. Annotating HS and stereotypes in the FB-Stereotypes corpus

In this Section, we mainly focus on the annotation scheme designed for FB-Stereotypes and its application on this corpus. The first subsection is especially devoted to the description of the labels used for the annotation of each dimension, provided together with the constraints we defined for them, while the second is about their application on data.

4.1. An annotation scheme for HS, stereotype and discredit

The main focus of the annotation is stereotype, but, as mentioned before some ancillary features can be useful for better describe the faceted nature of this phenomenon.

For what concerns the dimension of **HS** we followed the guidelines[16] provided for annotation the annotation of benchmark corpora used in the two editions of HaSpeeDe task (Sanguinetti et al., 2020). In order to be annotated with a value signaling the

[16] https://github.com/msang/hate-speech-corpus/blob/master/GUIDELINES.pdf

Table 2
Set of examples from the FB-Stereotypes dataset.

Type	Text	Translation
proself	certo che i nostri genitori anche i miei sono andati a lavorare all estero, ma legalmente e rispettando le leggi..	→ *sure that our parents also mine went to work abroad, but legally and respecting the laws..*
proself	con la differenza che gli italiani lavoravano non violentavano non portavano malattie e non rubavano le case del paese che li accoglieva, ma ve la date una svegliata	→ *with the difference that the Italians worked they did not rape they did not carry diseases and they did not steal the houses of the country that welcomed them, would you wake up*
prosocial	E' solo un caso. E dovremo essere grati a Dio, al destino, alla sorte, a qualsiasi cosa per il fatto di essere nati qui e non lì. Povera gente, quanta sofferenza!	→ *It's just a coincidence. And we should be grateful to God, destiny, fate, anything for being born here and not there. Poor people, how much suffering!*
prosocial	Non trovo differenze, in queste foto vedo la stessa disperazione.	→ *I find no differences, in these photos I see the same desperation.*

presence of the dimension "HS" a tweet must be a message that spreads, incites, promotes or justifies hatred or violence against the given target, or a message that aims at dehumanizing, delegitimating, hurting or intimidating the target. For instance, the tweet "*La prossima resistenza la dovremmo fare subito contro gli invasori islamici!*" (We should start fighting back Islamic invaders right now!) was annotated as HS, since it presents the members of a religious group as a dangerous enemy – which is harmful and delegitimating – and calls for a violent action against them.

Different strategies are used in literature for the annotation of HS, see e.g. Poletto, Basile, Bosco, Patti, and Stranisci (2019), each showing advantages and troubles. In this corpus, provided the difficulty and the complexity of evaluating the presence of HS tout court, we have moreover applied a scalar annotation rather than the binary one applied e.g. in Bosco et al. (2018) using values from 0 to −3 for HS, namely no HS, weak HS (−1), medium HS (−2) and strong HS (−3), whose distribution will be described in detail in Section 4.2.3. Following the hypothesis that the notion of **stereotype** can be especially relevant for investigating HS, we annotated this second dimension in a finer-grained way. First of all, we distinguished stereotype from the evaluative and affective dimension of **prejudice** (Brown, 2011); furthermore we were interested in detecting the evaluative dimension underlying both stereotype and prejudice in terms of socio-cognitive model of evaluation (Miceli & Castelfranchi, 2000), i.e. **discredit**. Therefore we annotate as a binary category stereotype (yes/no) and in cascade the binary category prejudice (yes/no) and the multivalued category discredit, whose values are better described below.

When we consider the evaluative process behind the stereotype or the prejudice, we can negatively evaluate and then spoil the other's image according to six main criteria useful for identifying the different forms that discredit can assume (D'Errico & Poggi, 2012; D'Errico, Poggi, & Vincze, 2012; Poggi, D'Errico, & Vincze, 2011) and also by taking into account the *Stereotype Content Model* (SCM) proposed by Fiske and colleagues (Fiske, Cuddy, & Glick, 2006), recently adapted to immigrants stereotypes in D'Errico et al. (2022). In our annotation scheme the discredit category can assume the following values:

- (Attack to) **Benevolence**: acting on behalf of their interests and not one's own, being trustworthy, honest, ethical. In this case the immigrant will be seen as a thief, a rapist, a murderer, a criminal and a profiteer: an unreliable individual that behaves illegally. Example: *"certo che i nostri genitori anche i miei sono andati a lavorare all estero, ma legalmente e rispettando le leggi"*. (sure that our parents also went to work abroad, but legally and respecting the laws ..).
- **Competence**: it emerges from the experience, knowledge and intelligence of an individual. On the competence dimension, an individual is negatively evaluated by accusing him of ignorance, stupidity, lack of preparation, incomprehensibility. The immigrant will thus be an imbecile, a goat. Example: *"Ma smettiamola di paragonare gli italiani a loro qua! Noi abbiam creato un sacco di cose! Loro cosa? Ma dai va la"* (But let's stop comparing the Italians to them here! We have created a lot of things! Them what? But come on).
- **Affective Competence**: competence can also be affective when the immigrant is described as a numb, emotionless person, or a parasitic, describing the individual's emotional features (Poggi & D'Errico, 2009). Example: *"come si comportano molti di loro,ci mancano di rispetto in tutti imodi"* (as many of them behave, we disrespect us in every way).
- **Dominance**: the person can be described in terms of power. This dimension is about having or not having power, the ability to influence others and impose one's will. In particular:

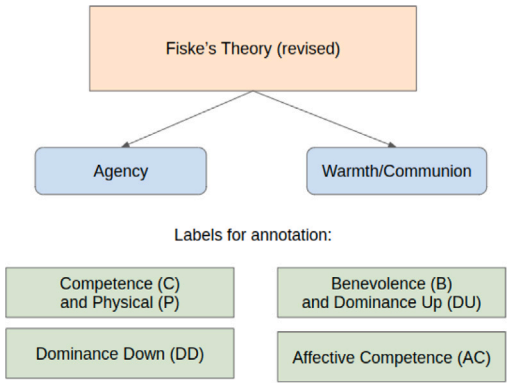

Fig. 2. Simplification of Fiske's SCM theory on agency and warmth.

- in passive terms (**Dominance Down**): immigrant can be described as a parasite, a nothing, a slacker who complains. Example: *"Caro Gianni i nostri emigrati andavano a lavorare non a farsi mantenere ....."*. (Dear Gianni, our emigrants went to work not to be maintained by us ......).
- in active terms (**Dominance Up**): when the immigrant is classified as overbearing, dangerous, aggressive, someone who also forcibly demands something he hasn't deserved. Example: *"Mio padre è stato un immigrato ma è andato a zappare la terra....ora non lo vuol fare più nessuno ...chissà come mai.... Loro vengono qui e vogliono il lusso...."*. (My father was an immigrant but went to hoe the earth .... now nobody wants to do it anymore ... who knows why ... They come here and want luxury .....).

- **Physical** Features: dirty, sick, ugly, disgusting, black, dangerous for the health of the citizens. Example: *"Ora arrivano qua in massa senza nessun permesso, asenza nessuna visita, hanno rioportato da noi malattie che avevamo combattuto e debellate da almeno 40 anni"*. (Now they arrive here en masse without any permission, without any visit, they have brought back diseases that we had fought and eradicated for at least 40 years.).

It should moreover be observed that according to the SCM these six forms of discredit can be grouped together in two main dimensions: *agency* and *warmth*, in some study also referred as *communion*. While agency refers indeed to individual qualities relevant for goal-attainment, such as being ambitious or capable or powerful, warmth refers to qualities relevant for the establishment and maintenance of social relationships, such as being friendly/kind/loving or fair. Agency and warmth, thus, seize the two major recurring challenges of human life: pursuing individual goals and belonging to social groups (Ybarra et al., 2008).

Within the dimension of *agency* Competence (**C**), Dominance Down (**DD**) and Physical (**P**) can be included. At the other end of the spectrum, *warmth* may comprise Benevolence (**B**), Dominance Up (**DU**) and Affective Competence (**AC**), moving from a finer-grained to a rougher-grained annotation.

The granularity of the annotation of discredit can be also simplified by organizing the six original forms of discredit in four dimensions, rather than in two, grouping together Competence and Physical, Benevolence and Dominance Up, but keeping separate the other two forms (i.e. Affective Competence and Dominance Down) as shown in Fig. 2. This solution of compromise between the finer granularity, based on six categories, and the rougher one, based on two only, may be also suggested by the distribution of the categories of discredit in the dataset, where a recurrent co-occurrence of Benevolence and Dominance Up or Competence and Physical feature has been observed. All the three granularity are observed and discussed in the analyses and experiments provided in the next sections.

To summarize, in the schema, stereotype has been considered as an independent component with respect to HS, so that the latter does not entail the former, and viceversa. Nevertheless, to be compliant with the definition in literature of the phenomena we are annotating, we have imposed some constraint in the annotation of the other dimensions. In particular, coherently with Brown's theories (Brown, 2011) and with other literature (Warner & Hirschberg, 2012), prejudice has been deemed an explicit form of stereotype and thus bound to the latter's presence: in the annotation process there could be stereotype without prejudice, but less likely prejudice without stereotype. Moreover, only if prejudice was present, annotators were asked to clarify the evaluative dimension of prejudice in that particular comment and to subsequently annotate a label among those provided for the discredit dimension. A further constraint is to ask the annotator to select a single label for discredit, selecting the prevailing one in case where they perceive the occurrence in the message of more than one form of discredit.

The categories annotated in FB-Stereotypes and the corresponding labels used for marking the comments on immigrants are therefore as follows:

- HS: 0, −1, −2, −3
- stereotype: yes, no
- prejudice: yes, no
- discredit: benevolence = B, competence = C, affective competence = AC, dominance up = DU, dominance down = DD, physical feature = P.

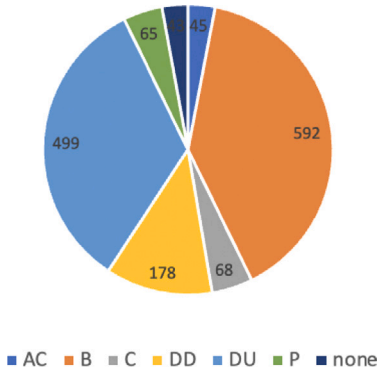

Fig. 3. Distribution in absolute values of the category discredit in the proself portion of the FB-Stereotypes corpus, which correspond to the following percentages: B 40%, DU 33%, DD 12%, C 5%, P 4%, AC 3%.

4.2. Application of the scheme on the FB-Stereotypes corpus

As far as the annotation process is concerned, four annotators have been involved, two computational linguists, one social psychologist and one master degree student skilled in linguistics.[17] Three among them independently applied the annotation on the whole proself portion of the corpus, originally composed of 1,504 comments, but cleared of 14 message discarded because at least two annotators indicated that their content was not comprehensible or out of topic. On the dataset composed by 1,490 posts the annotators achieved the agreement for a large portion of the data. The fourth annotator provided a further annotation only for the cases where disagreement among the three other ones has been detected during the disagreement analysis.

4.2.1. Annotation of stereotype and prejudice

For what concerns the dimension of stereotype, which is the main focus of our annotation, the annotators achieved a full agreement selecting the same value in a larger portion of messages with respect to that achieved for HS (see Section 4.2.3), being only 100 over 1,490 the posts annotated by using different labels by the different annotators.

The Fleiss's kappa calculated for the category stereotype over the annotation of three human judges is $\kappa = 0.558$, while for the category prejudice is a little bit lower, i.e. $\kappa = 0.475$. A larger amount of messages where the annotators did not achieve the agreement (193) has been found for what concerns the dimension of prejudice. Considering that the prejudice has been assumed, according to the definition in psychological literature (Brown, 2011; Warner & Hirschberg, 2012), as dependent on stereotype. Around one half of the cases in disagreement for prejudice are those featured by disagreement also for stereotype as cited above.

4.2.2. Annotation of discredit

For the dimension of discredit, which presents a greater number of labels among which the annotators are asked to select the more suitable one, the full agreement among the annotators has been achieved in around two thirds of the corpus (1,014 posts). The distribution of the labels shows that the label B (Benevolence) is the most often annotated (in 500 posts), followed by DU (Dominance Up, in 335 posts), while a few cases were annotated with the other ones and more precisely 71 DD (Dominance Down), 43 P (Physical feature), 33 AC (Affective Competence), 32 C (Competence). Fig. 3 shows percentages for this label distribution. The inter-annotator agreement was also calculated on the different forms of discredit but due to the unbalanced distribution of labels and scarcity of data the result was inconclusive.

In the first stage of development of the corpus here described, we asked the annotators to select only one label to annotate the discredit in each message, as written above about the constraints applied in the annotation. Nevertheless, in some messages of our corpus, which are drawn from Facebook and therefore have no limits in length (as it happens on other micro-blogging platforms such as Twitter, for instance), more than one form of discredit and therefore more than one suitable label for the annotation of this dimension may occur.

Future developments of this work will include experimenting an annotation where multiple labels can be selected for a single message, in order to obtain a more faceted and more faithful representation of the content.

4.2.3. Annotation of hate speech

Our main focus in the development of the corpus is on stereotypes. All data included in the proself portion of FB-Stereotypes are selected because they express opinions against hosting immigrants and therefore our expectation was that almost all were also including HS towards them. The distribution of HS confirms this hypothesis.

[17] For a detailed document of guidelines regarding the annotation of HS, Stereotype and discredit, please refer to: https://github.com/boscoc/stereotypes/find/main.

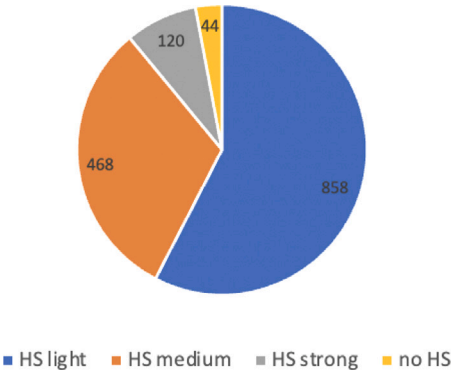

Fig. 4. Distribution in absolute values of the category HS in the proself portion of the FB-Stereotypes corpus, which in percentage corresponds to HS weak 58%, HS medium 31%, HS strong 8%, no HS 3%.

The distribution of the labels shows that users expressed in a quite polite form their opinions, see absolute values and percentages in Fig. 4 and its caption; HS = −1 is indeed the label selected for more than one half of the messages (858 posts corresponding to 58%). A relevant amount of users expressed more aggressive and offensive opinions, i.e. marking HS = −2 (468 posts corresponding to 31%), while only a few used very strong HS messages, i.e. marking HS = −3 (120 posts corresponding to 8%). The rest of messages (44 corresponding to 3%) have been marked with HS = 0, i.e. not HS against the target.

In the FB-Stereotypes corpus in the annotation of this dimension the annotators achieved a full agreement selecting the same value on the scale (from 0 to −3) in the large majority of posts (about 1,245). This agreement is higher than that reported in other cases of annotation of HS corpora (Poletto et al., 2017) and we can hypothesize that this outcome is influenced by the simultaneous annotation of stereotype that binds the annotators to reflect more thoroughly on HS too.

As far as the other section of the corpus is concerned, that where the messages are *prosocial* i.e. express attitudes favorable to the immigrants and refugees, as said above, some trial of annotation has been done for determining if the stereotype can occur or not in these messages. We annotated 200 messages randomly selected and we have validated the hypothesis that racial stereotype does not co-occurs with prosocial attitude. If some stereotype occurs in this portion of data it could be only referred to other category of people. See below some examples:

Ex.1 Grande Gianni e grazie che hai ricordato a tutti che anche i nostri nonni e bisnonni hanno fatto la stessa cosa degli attuali migranti in cerca di un futuro migliore.
→ *Awesome Gianni and thanks for reminding everyone that also our great-grandparents did the same thing of the current migrants in seek of a better future.*

Ex.2 Infatti Gianni hai ragione, ma la gente ha la memoria corta
→ *Indeed gianni you're right, but people have short memory*

Ex.3 povera gente, alla ricerca della liberta' e di un mondo migliore.Poveri bambini.....
→ *poor people, in pursuit of freedom and of a better world.Poor children.....*

If some form of HS occurs, it is not against our reference target, namely immigrants. This allowed us to consider this portion of data as representative of non stereotype and non HS against immigrants. The usefulness of this portion of data will be made clear in the next section, where we describe the computational experiments.

5. Lexical analysis and discussion

In this section, we present a detailed lexical analysis that will also provide some hints about further investigation. Moreover, we will exploit in this analysis small sets of data extracted from another corpus which has been used as benchmark in an evaluation campaign for the detection of HS (HaSpeeDe2020).

In order to better understand the linguistic characteristics of the dataset taken into account in the present study, we performed the following lexical analyses:

i. listing the most relevant n-grams (unigrams, bigrams and trigrams) of the *proself* dataset, and comparing them with the respective n-grams of the *prosocial* part;
ii. listing of the most relevant n-grams of the datasets annotated with stereotype = *yes*, and comparing them with the n-grams of the datasets annotated with stereotype = *no*;
iii. listing the most relevant n-grams from all the datasets with respect to the labels of discredit (AC, B, C, DD, DU, P).

Table 3
Discriminating lexica according to proself or prosocial type.

prosocial lexica	TF-IDF	proself lexica	TF-IDF
memoria (*memory*)	79.71	albergo (*hotel*)	87.26
ignoranza (*ignorance*)	68.56	pretendere (*to claim*)	76.35
umanità (*humankind*)	57.22	cazzata (*bullshit*)	62.63
corto (*short*)	50.54	regolare (*regular*)	61.47
abbracciare (*to hug*)	47.59	buonismo (*do-goodism*)	61.44
memoria corto (*short memory*)	45.04	quarantena (*quarantine*)	61.36
complimento (*compliment*)	45.00	culo (*ass/luck*)	60.51
esportare (*to export*)	41.02	paragonare (*to compare*)	58.65
signora (*madam*)	39.37	alloggiare (*to accomodate*)	55.64
odiare (*to hate*)	38.83	mantenuto (*economically supported*)	54.21

Before performing the linguistic explorations accordingly to the three settings just outlined, all texts have been pre-processed by deleting all user mentions, stop-words, URLs and by leaving only words that were lexically significant. Furthermore, they have been tokenized, lemmatized and normalized to lowercase letters with the SpaCy library.[18] Also, all punctuation has been stripped.

In the following subsections are reported the outcomes of the three main lexical analyses in detail.

5.1. Proself versus prosocial

The first exploration we wanted to perform was to investigate whether there was any kind of lexical divergence between the two main types of data: *proself* and *prosocial*. In order to do so, we compared all the messages contained in the whole original corpus composed of 12,095 cited before, as seen in Table 1.

After having pre-processed the messages, we extracted n-grams and computed the Term Frequency-Inverse Document Frequency (TF-IDF) in order to observe which are the words that have the highest relevance in the two portions of the dataset. In Table 3 we report the 10 highest ranking words with the respective cumulative TF-IDF value obtained by summing all the single scores for each message contained in the dataset (because of this, the values of TF-IDFs range from 0 to 100). The samples of words belonging to the two different portions of the dataset seem to be lexically disjointed.

Furthermore, extending the view on a larger set of words occurring the two portions of the dataset, we can observe the distribution of verbs on the one hand and noun+adjectives on the other hand with respect the two types of texts, and that they are featured by a further meaningful lexical difference.

Proself

VERBS: andare lavorare (*to go to work*), rubare (*to steal*), mantenere (*to keep*), cantare (*to sing*), venire (*to come*), paragonare (*to compare*), emigrare (*to emigrate*), arrivare (*to arrive*), dovere (*to must*), cercare (*to search*), delinquere (*to commit a crime*), pretendere (*to claim*).

NOUNS+ADJECTIVES: italiano (*Italian*), vero (*true*), casa (*home*), paese (*country*), italia (*Italy*), differenza (*difference*), paragone (*comparison*), albergo (*hotel*), soldo/i (*money*), buonismo (*do-goodism*), diverso (*different*), povero (*poor*), problema (*trouble*), regola/e (*rule/s*).

Prosocial

VERBS: stimare (*to respect*), pensare (*to think*), dovere (*must*), sapere (*to know*), ricordare (*to remember*), condividere (*to share*), sentire (*to feel*), scappare (*to flee*), ignorare (*to ignore*), capire (*to understand*), abbracciare (*to hug*), bisognare (*to need*).

NOUNS+ADJECTIVES: commento (*comment*), umano (*human*), memoria (memory), persona (*person*), ragione (*reason*), storia (*history*), ignoranza (*ignorance*), cuore (*heart*), razzista (*racist*), umanità (*humankind*), morto/i (*dead/s*), pensiero (*thought*).

From a first glimpse at the lexical entries it is clear that in the proself part of the data, most of the words used belong to a semantic sphere that has a negative polarity (e.g., steal, committing crimes, demand, differences, different, poor, problems). On the other end of the spectrum, in the prosocial part of the data, the most interesting words seem to have a positive polarity and the words have a connection with the affective sphere (e.g., respect, feel, hug, need, heart) and the cognitive sphere (e.g. think, know, ignore, understand, memory, history, ignorance, thought).

5.1.1. Stereotype = yes versus Stereotype = no

In this second part of the analyses we want to investigate the lexical differences of messages in relation to the presence/absence of stereotype. In order to do so, we joined two different datasets that are both annotated with labels that describe stereotypes, and also belong to different textual genres. Firstly, we take into account the FB-Stereotypes dataset described in the sections above, and secondly we merge it with a portion of 201 tweets extracted from the HaSpeeDe2020 corpus that have also been annotated with the same labels (in particular: *stereotype = yes/no*).

Hence, for the data that is annotated as *stereotype = yes*, we have considered all the Facebook messages of the proself part that contain stereotype and an addition of 98 tweets from the HaSpeeDe2020 corpus. While for the data that is annotated as *stereotype*

[18] https://spacy.io/

Table 4
Discriminating lexica according to presence or absence of stereotype.

Stereo = yes	TF-IDF	Stereo = no	TF-IDF
pretendere (*to claim*)	39.33	commento (comment)	27.99
preteso (*claimed*)	20.34	memoria (*memory*)	24.00
mantenuto (*kept*)	18.35	corto (*short*)	17.09
cercare lavorare (*to try to work*)	17.37	de (*of*)	16.99
paragonare (*to compare*)	17.04	ignoranza (*ignorance*)	15.19
comportare (*to behave*)	16.97	condividere (*to share*)	14.97
regolare (*regular*)	14.95	memoria corto (*memory short*)	14.94
gratis (*free*)	14.35	ragione gianni (*reason gianni*)	12.33
invasione (*invasion*)	14.29	umanità (*humankind*)	12.00
farsi mantenere (*to be maintained*)	13.99	razzismo (*racism*)	10.28

= *no*, we have considered all the Facebook messages of the prosocial part and the addition of 103 tweets from the HaSpeeDe2020 corpus.

Also in this scenario the texts have been cleaned, tokenized and lemmatized. Afterwards, the TF-IDF measure was calculated on the two different portions and the 10 highest ranking words for both classes are reported in Table 4.

Once again, we report the values of the cumulative TF-IDF obtained by summing all the single scores for each message contained in the dataset. Interestingly we observe the presence of weird unigrams and bigrams such as "of" which can be imputed to errors of automatic lemmatization.

However, interesting findings are expressions such as "memory short" or "to go to work" that seem to suggest some sort of lexical relevance with respect of the phenomenon investigated here.

5.1.2. Forms of discredit

In this third and last section of the lexical analyses, we wanted to investigate the linguistic relevance according to the different categories of discredit. According to the annotation scheme that has been followed, the forms of discredit can be annotated only when stereotype is present. Due to this design, for this step, we rely on the messages used in the precedent step (where *stereotype = yes*).

In the same way as we did in the two prior steps, also here the texts have been pre-processed and the TF-IDF measure was calculated in order to understand which words are the most relevant for each form of discredit. Following, a non-exhaustive list is presented:

- AC = buttare (*to throw away*), educare (*to educate*), comportare civile (*to respectufully behave*), differenza (*difference*), cultura (*culture*), lavoratore (*worker*), ospitare (*to host*), potere (*to can*);
- B = rubare (*to steal*), clandestino (*illegal*), delinquere (*to commit a crime*), differenza (*difference*), rispettare (*to respect*);
- C = differenza (*difference*), diverso (*different*), cultura (*culture*), costruire (*to build*), paragonare (*to compare*), comportare (*to behave*);
- DD = andare a lavorare (*to go to work*), parassita (*parasite*), cercare (*to look for*), volere (*to want*), mantenuto (*economically sustained*), soldi (*money*);
- DU = pretendere (*to claim*), venire (*to come*), casa (*home*), preteso (*claimed*), mantenere (*to economically sustain*), andare a lavorare (*to go to work*);
- P = malattia (*sickness*), bello (*beautiful*), feccia (*excrement*), confondere (*to confuse*), quarantena (*quarantine*), scarafaggio (*coakroach*).

Taking into consideration the fact that the six categories of discredit present a highly unbalanced distribution in the whole corpus (as the majority of texts belongs to the categories B, DD and DU, while AC, C and P contain very few instances, see also Fig. 3) it is not feasible to derive any consistent assumption. Any linguistic intuition we might derive from this lexical analysis might not be perfectly adherent to reality and should be consolidated in a future, larger version of the corpus where more data will be annotated accordingly to the same scheme (see Future Work).

6. Experiments

In this section, we aim at shedding some light on the detection of stereotypes by designing it as a binary classification task, that is the identification of messages where some stereotype occurs versus those where it does not. Therefore, we have performed experiments where we train on the dataset of comments from FaceBook described above (FB-Stereotypes) and a portion of a benchmark dataset of tweets and news headlines (HaSpeeDe2020) some classifiers whose performance has been well consolidated in classification tasks.

The datasets taken into consideration for the experiments presented in this section, are reported in the following Table 5.

In the evaluation phase, we carried out a testing on the tweets coming from a novel corpus soon-to-be released, which is composed of tweets including racial hoaxes about immigrants (Reactions-to-Hoaxes). This new set of data represent in some more general sense the discussion about immigrants, which is different from the FB-Stereotypes focused on the specific case raised by Gianni Morandi.

Table 5
The distribution of training and test set.

Set	Dataset	Total
training	FB-Stereotypes	2,990
training	HaSpeeDe2020	201
test	Reactions-to-Hoaxes	199

Table 6
Results of the baselines for the detection of stereotypes. *(no = negative class, yes = some stereotype occurs in the message).*

Stereotype			
SVM-Model	`no`	`yes`	`w_avg`
precision	0.629	0.714	0.664
recall	0.949	0.185	0.638
F1 score	0.757	0.294	**0.569**
GilBERTo-Model			
precision	0.671	0.760	0.705
recall	0.949	0.257	0.682
F1 score	0.786	0.384	**0.631**

Moreover, in order to understand the difference of the main topics between the training and the test set, we performed a 10-fold cross-validation on the texts of the training set.

In particular, we designed different models exploiting classical and deep learning algorithms, with the aim of understanding the contribution of the discriminative lexicon presented in Section 5 in the automatic detection of stereotypes. We employed the Support Vector Machine (SVM) algorithm, which has been previously successfully used in text classification tasks, especially for detecting hate speech (Fortuna & Nunes, 2018), and we also experimented fine-tuning the GilBERTo language model enhancing the knowledge of the system with a model pre-trained on a big amount of Italian data. This new approach revealed to reach impressive scores in various NLP related tasks (Qiu et al., 2020), as well as in stereotype detection. Indeed, as shown in the ranking reported on Twitter data in Sanguinetti et al. (2020), the best system in stereotypes detection used a simple fine-tuning of an Italian pretrained LM.

6.1. Computing baseline measures for stereotype detection

Firstly, we designed two baseline models that could help us to measure the contribution of the lexical information in a classifier of stereotypes online (Section 6.2):

- the SVM with the Radial Basis Function (RBF) kernel[19] and the default parameters provided by scikit-learn library[20] trained only on texts represented as a vector of frequencies taking into account the 500 most frequent unigrams (SVM-Model);
- the fine-tuning of GilBERTo language model (Ravasio & Di Perna, 2020) on stereotypes detection, taking into account only the CLS token[21] of the BERT-based model (GilBERTo-Model). Indeed, in accordance with Devlin, Chang, Lee, and Toutanova (2019), the purpose of this token is to contain the information useful for the classification task at the end of the forwarding process. Then a simple classifier can just take this CLS token as input to classify the whole text. Moreover, we added a dropout layer and a final linear layer to get the class-related probability employing a Sigmoid function.[22]

The evaluation metrics used in these experiments are the common ones used to validate the models of text classification: precision, recall and F1 score. These measures are calculated on each class (`yes` and `no`) and on the weighted average of both classes (`w_avg`). In Table 6, we report the results obtained by the selected baseline models on the binary classification task of stereotype detection.

From the observation of Table 6 it can be seen how the F1 score obtained with GilBERTo is higher than the one obtained by using a plain SVM (0.631 vs. 0.569). These results confirm the utility of the pretrained LMs respect to the employing of classical machine learning, as already shown in the results obtained in the second edition of the HaSpeeDe competition in tweets.

Furthermore, with both systems, the negative class seems to be captured meaningfully better than the positive one. Considering that the most desirable outcome for a detection tool is the correct recognition of the observed phenomenon (in our case, the presence of stereotype), this is not an ideal outcome of the classifier. Due to this reason, in the next section we provide further experiments in which we enhanced our models by lexical knowledge related to stereotypes.

[19] Explorative experiments attested the usefulness of RBF kernel.

[20] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[21] Classification token in BERT.

[22] We only report here the scores achieved by using GilBERTo that is the best performing language model; while the experiments we performed with AlBERTo (Polignano, Basile, Basile, de Gemmis, & Semeraro, 2019) and UmBERTo https://github.com/musixmatchresearch/umberto, which produced less interesting scores, are not reported.

Table 7
Results of the lexically enhanced models for the detection of stereotypes. *(no = negative class, yes = positive class).*

Stereotype			
SVM_lex-Model	no	yes	w_avg
precision	0.640	0.551	0.604
recall	0.813	0.333	0.618
F1 score	0.716	0.415	**0.594**
GilBERTo_lex-Model			
precision	0.697	0.620	0.667
recall	0.839	0.419	0.677
F1 score	0.761	0.500	**0.661**

6.2. Computing lexically informed approaches for stereotype detection

In this section, we present the experiments performed informing the previous systems with lexical information provided by the lexicon seen in Section 5. In particular, our models have been designed as follows.

- The SVM-based system takes as input the texts represented with the normalized values of the TF-IDF measure for each word existing in the discriminative lexicon (SVM_lex-Model).
- The main idea beside the lexical informed GilBERTo-based model is to converge the awareness coming from a pre-trained language model with the specific knowledge derived from dedicated lexical features. On the one hand, the learning transferred by a language model trained on different Italian texts coming from the web should help the classifier to generalize better, ranging from informal and more formal writings, and should make the system able to 'understand' better the unseen texts. On the other hand, lexical features lead the system to pay attention to specific elements in the text. Therefore, we added at the previous network of GilBERTo-Model a new input layer consisting in the batch normalization of the features vector (GilBERTo_lex-Model),[23] combined with CLS token of BERT-based pre-trained model.

Table 7 shows the scores achieved for these settings, from which two main findings arise. The first one is that a general very low improvement of scores is achieved by adding lexical information to both models employed; indeed the weighted averaged F1 score of SVM improves from 0.569 to 0.594 ($\Delta = 0.04$), and that of GilBERTo grows from 0.631 to 0.661 ($\Delta = 0.05$). The second finding is instead that, for both models, the addition of lexical information determines a meaningful reduction of the difference between the F1 score for the positive class and that for the negative class (with respect to the results reported in Table 6), showing a better recall of positive instances (texts containing stereotypes).

On the one hand, this leads us to formulate the hypothesis that the discriminative lexicon extracted by linguistic relevance (TF-IDF) to the category of stereotype is of some help for capturing such phenomenon. Nevertheless, the general poor contribution of lexical information also leads us to conduct further analyses about: (1) how the weights of words from the lexicon are really computed inside our classifier, and (2) the nature of our training and test set.

6.3. Further analyses

Taking into account the aims of our further analyses, firstly, we examined the weights computed inside the SVM_lex-Model during the testing; and then, we performed a 10-fold cross validation on the training set (shuffling the data) to understand the impact of lexical features in a corpus that contains especially comments about a singular event (the post of Gianni Morandi).

For the first analysis, we exploited the SHAP library.[24] that give us the possibility to visualize the relevance of the words in the lexicon for the classification process.

Fig. 5 shows the 20 most relevant words in the task of classification with the SVM_lex system. In particular, we can notice that words with negative values (in blue) help the classifier to detect the negative instances (the texts without stereotypes), whereas the positive values (in red) help the classifier to detect the positive instances (the texts with stereotypes).

Looking at this figure, we can notice that the words especially relevant for the positive class are actually less intuitive. However, looking at the test set, we observed that words such as "domenica" (*Sunday*) or "vittima" (*victim*) are present in texts where the users tend to express their frustration about the perceived difference between the 'favorable' condition of actual immigrants in Italy and the past condition of Italian immigrants in other countries, like:

> *caro Gianni gli italiani venivano sfruttati [...] sappi che io ho mio padre che percepisce una pensione di 480 euro mensili lordi, dopo aver lavorato dall'età di 6 anni fino a 79 anni senza mai fare una vacanza e lavorando pure la domenica, insieme a mia madre*

[23] The batch normalization technique helps to standardize the layer and stabilize the learning process.
[24] https://shap.readthedocs.io/en/latest/index.html

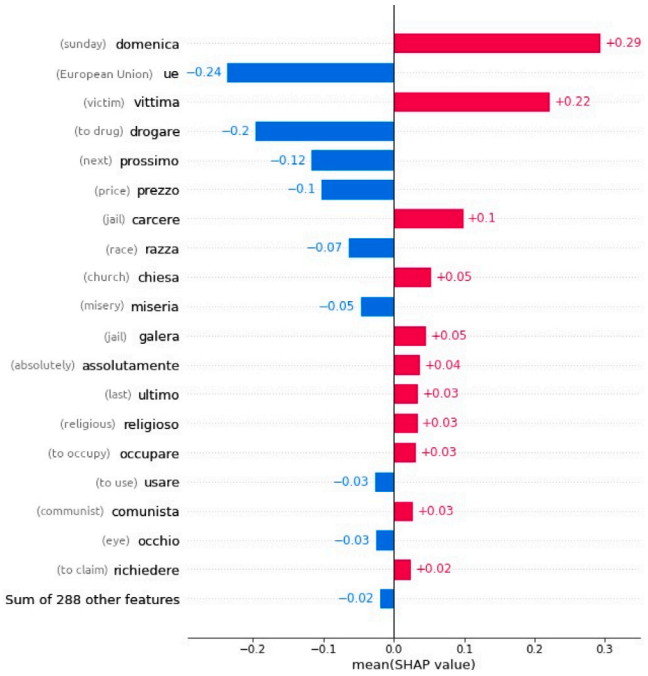

Fig. 5. Local feature importance plot.

ormai deceduta. come vedi hanno molti più diritti room ed extracomunitari rispetto a chi a lavorato tutta una vita.. fosse per me li rimanderei tutti nei suoi paesi compresi tutti coloro che riempiono le carceri. [...]

→ Dear Gianni, the Italians have been exploited [...] you have to know that my father who receives a pension of 480 euros as gross per month, after having worked from the age of 6 to 79 without ever having holiday and also working on Sundays, together to my mother now deceased. as you can see, they, Romas and non-EU, have many more rights than those who have worked all their life .. for me, I will send them all back to their countries, including all those who fill the prisons. [...]

Even on the opposite, relevant but not intuitive words are used in specific contexts that make them important for the recognition of the messages without racial stereotypes. The most curious is the word "UE" (European Union) that we found in texts like:

Ma quanta disinformazione! Ai migranti non vengono dati affatto 30 euro al giorno! I soldi sono stanziati dalla UE per le strutture che li ospitano. Mantenuti? Ma lo sapete che per un richiedente asilo è VIETATO lavorare, per legge??? E l'iter dura almeno due anni! »un miracolo se non diventano delinquenti o muoiono di fame. Questa è la vergogna. Bravo Gianni!

→ But how much misinformation! Migrants do not receive 30 euros a day at all! The money is allocated by the EU for the structures that host them. Maintained? But did you know that for an asylum seeker it is FORBIDDEN to work by law??? And the process lasts at least two years! It's a miracle if they don't become delinquents or starve. This is the shame. Bravo Gianni!

In Fig. 6 we instead display the global feature importance plot, where the global importance of each word is taken to be the mean absolute value for that word over all the given samples.

Finally, to investigate the nature of our training and test set we applied a 10-fold cross-validation strategy. The effect of training and testing both on data from the same set containing texts mainly focused on a specific event (the post of Gianni Morandi), consists in neutralizing the impact on scores of any linguistic (also lexical) difference between the training corpus and that of testing. When a model is tested by using this strategy, the results are expected as better than in the evaluation on different set of data. The scores we achieved, reported in Table 8, clearly confirm this trend. The table indeed shows a meaningful improvement of scores with respect to the F1 scores reported above (Tables 6 and 7) for what concerns the average and each single class identification.

More precisely, the upper part of Table 8 shows more details about these expected improvement, since the weighted average of the F1 score of SVM-Model increases by almost a Δ of 0.3 with respect to the value reported in Table 6 for the same setting but performing on a different test set. An also higher improvement may be referred to the positive class, which score goes from 0.294 to 0.673.

A similar situation can be observed also in the lower part of Table 8, that gives instead some hints about the contribution of lexical knowledge. In the lexically enhanced setting SVM_lex-Model indeed improves the `w_avg` of almost a Δ of 0.3 with respect to the value reported in Table 7 for the evaluation on the reactions to racial hoaxes. And also in this case the higher improvement may be referred to the positive class, which score goes from 0.415 to 0.724.

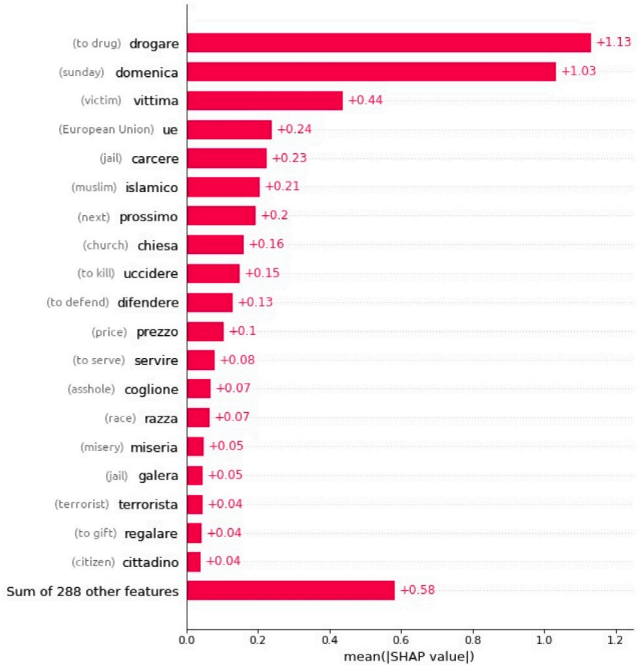

Fig. 6. Global feature importance plot.

Table 8
Results of the 10-fold cross-validation of the SVM-based classifiers without lexical enhancement and with it for the detection of stereotypes. *(no = negative class, yes = positive class).*

Stereotype			
SVM-Model	no	yes	w_avg
F1 score	0.787	0.673	0.732
SVM_lex-Model			
F1 score	0.795	0.724	0.761

Therefore, in general we noticed that the introduction of lexical information specific for stereotypes against immigrants helps the system, regardless the adopted algorithm of learning, to detect the presence of these stereotypes in the messages online respect for instance to the simple fine-tuning. However, this lexicon seems to be dependent on the context of reaction to the Gianni Morandi's post.

7. Conclusion and future work

This paper presents and discusses an annotation scheme for making explicit the presence of stereotype and some related phenomena, such as HS and discredit. Stereotype is indeed the cognitive nucleus of prejudice which assumes in turn the face of discrimination and hateful behavior. Provided the relevance in the last few years of HS detection and other alike tasks, e.g. cyberbullying detection, offensive language and misogyny identification, this paper aims at shedding some novel light on HS by investigating stereotype for paving the way for the improvement of linguistic resources and tools for HS detection.

The paper starts with a brief survey of the notions of HS, stereotype and prejudice (Sections 1 and 2), which are spreading in social media and newspapers and are among the hottest topics for text classification and computational linguistics. Then, to better analyze these notions, we introduce a case study: FB-Stereotypes that is a corpus for Italian drawn from Facebook including texts about immigrants where the annotation scheme has been applied (Section 3). The corpus includes a balanced amount of messages against hosting immigrants and in favor of hosting them in Italy (namely proself and prosocial). The preliminary analysis of this corpus shows how HS and stereotype co-occur. It moreover allows us to observe a higher agreement in the annotation of the latter phenomenon, according to the idea known in the literature that stereotype is a social construct (Section 4).

Benefiting from the strategy applied for the collection of data, we have included in this corpus a percentage of hateful and stereotyped comments which is very high also concerning other corpora about immigration and related topics. The corpus collects indeed the comments to a message posted by a famous Italian singer that expressed empathy towards the victims of a shipwreck also comparing them with Italians who had emigrated to the USA in the 20th century. The content of the message has caused the

sending of several comments full of stereotypes making the corpus an especially interesting sample of data about the stereotypes that Italian have towards immigrants.

It is also interesting to note that Gianni Morandi's profile is not a far-right political one nor a populist page, and therefore, it does not naturally collect a population of followers with fixed skewed political ideas, but potentially a varied population by gender, origin and political ideas, homogeneous perhaps only by age (it should be verified). Yet, in the comments below that post, we encountered a great number of negative stereotypes and prejudices about immigration. This might be a clue about the different behavior of people based on their beliefs on the subject: "*those who hate to speak, those who do not hate are silent*". Indeed, even if the initial post is pro-migrant, even if the post is from a public figure that followers supposedly value and respect (given that they follow him on social media). Those who are against the topic feel entitled to criticize, and those who are in favor do not feel entitled to support their opinions enough.

Moreover, in Section 5 we carried out a lexical analysis of the corpus, whose most insightful outcomes are discussed in the paper. They show that almost all the categories we annotated in the corpus (at least those that are enough represented in it) are well characterized from the lexical point of view and can be therefore discriminated based on the lexical items they contain.

Finally, for validating the schema and the hypothesis that lexical features can be useful for the classification of racial stereotypes, we performed some classification experiments (Section 6). These preliminary experiments showed that the introduction of lexical information about racial stereotypes helps the system to detect their presence in messages online. However, the created lexicon seems to be dependent on the context of reaction to Gianni Morandi's post, probably for the big amount of data regarding this topic. Therefore, in future work, we want to extend the data from which we extracted the lexicon, including other reactions to other posts or tweets, to make this lexicon of racial stereotypes more specific and useful for real-world applications. In particular, we are currently working on the collection of a corpus of *"reactions to racial hoaxes"*, i.e., tweets that reply to racial hoaxes spread online. This idea is justified by the fact that we would like to extend the data with the same annotation scheme on the Twitter domain, besides the Facebook domain that we already have and, at the same time, investigate the interactions and the propagation of racial stereotypes in the conversational thread. Indeed, we are not only collecting tweets that reply to the main post (which in our case is always a racial hoax), but also replies-of-replies and so on.

Surely, the collection of such data has a twofold goal: it allows us to apply the annotation scheme to a bigger quantity of data, helping us in the validation process of the annotation scheme, and also it will help us in testing it on a different textual domain. Hopefully, with this data expansion, also the categories that are now under-represented will increase their number.

Furthermore, in future work, we plan to assess the validity of the annotation scheme by training a greater number of skilled annotators that will join the "STERHEOTYPES - Studying European Racial Hoaxes and Stereotypes" project in the forthcoming months.

Finally, we also plan to exploit the structure of Twitter's conversational threads to study the possible correlations with the phenomenon of stance. As a hypothesis, we think there might be a co-occurrence or some level of orthogonality between the opinion of a user (their support or being against a previous post) and the propagation of racial stereotypes in conversations. For this, we plan to use the SDQC scheme, as it has been exploited in prior work studying Twitter conversational threads and the spread of fake news and rumors (Gorrell et al., 2019).

CRediT authorship contribution statement

Cristina Bosco: Conceptualization, Methodology, Data curation, Writing – original draft, Supervision. **Viviana Patti:** Conceptualization, Methodology, Data curation, Writing – original draft. **Simona Frenda:** Software, Formal analysis, Writing – original draft. **Alessandra Teresa Cignarella:** Data curation, Investigation, Writing – review & editing. **Marinella Paciello:** Data curation, Writing – review & editing. **Francesca D'Errico:** Conceptualization, Data curation, Project administration, Funding acquisition.

Data availability

Data will be made available on request.

Acknowledgments

The work of all the authors is supported by the international project *STERHEOTYPES - Studying European Racial Hoaxes and Stereotypes* funded by VolksWagen Stiftung/Compagnia di San Paolo, Italy for the call for projects 'Challenges for Europe", (CUP: B99C20000640007); https://www.irit.fr/sterheotypes/.

Appendix

In Table 9, all the datasets mentioned in this work are summarized with the respective categories that are annotated within.

Table 9
Summary of all the different datasets used in this work and the categories they have been annotated for.

	Genre	Total_size	Annotated_size	HS
Gianni_proself	Facebook posts	7,046	1,490	−3, −2, −1, 0
Gianni_prosocial	Facebook posts	5,049	1,500	−3, −2, −1, 0
HaSpeeDe2020	tweets	8,102	50	yes/no
HaSpeeDe2020	news headlines	500	50	yes/no
Reactions_Hoaxes	tweets	5,255	199	N/A

	Stereotype	Context	Implicitness	Prejudice	Discredit
Gianni_proself	yes/no	N/A	N/A	yes/no	AC, B, C, DD, DU, P
Gianni_prosocial	yes/no	N/A	N/A	yes/no	AC, B, C, DD, DU, P
HaSpeeDe2020	yes/no	N/A	implicit/explicit	yes/no	AC, B, C, DD, DU, P
HaSpeeDe2020	yes/no	N/A	implicit/explicit	yes/no	AC, B, C, DD, DU, P
Reactions_Hoaxes	yes/no	yes/no	implicit/explicit	N/A	AC, B, C, DD, DU, P

	Stance	Aggressiveness	Offensiveness	Irony
Gianni_proself	N/A	weak/strong/none	weak/strong/none	yes/no
Gianni_prosocial	N/A	weak/strong/none	weak/strong/none	yes/no
HaSpeeDe2020	N/A	N/A	N/A	N/A
HaSpeeDe2020	N/A	N/A	N/A	N/A
Reactions_Hoaxes	S, D, Q, C	N/A	N/A	N/A

References

Allport, G. (1954). *The nature of prejudice*. Routledge.

Álvarez Carmona, M. Á., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H. J., Pineda, L. V., Reyes-Meza, V., et al. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J. C. de Albornoz (Eds.), *CEUR workshop proceedings*: *vol. 2150*, *Proceedings of the third workshop on evaluation of human language technologies for Iberian languages co-located with 34th conference of the Spanish society for natural language processing* (pp. 74–96). CEUR-WS.

Aronson, E., Wilson, T. D., & Akert, R. M. (2013). *Social psychology* (8th ed.). Pearson Education Inc.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 54–63).

Bauwelinck, N., & Lefever, E. (2019). Measuring the impact of sentiment for hate speech detection on twitter. In *Proceedings of HUSO 2019, the fifth international conference on human and social analytics* (pp. 17–22).

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018). Overview of the EVALITA 2018 hate speech detection task. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Eds.), *CEUR workshop proceedings*: *vol. 2263*, *Proceedings of the sixth evaluation campaign of natural language processing and speech tools for Italian. Final workshop co-located with the fifth Italian conference on computational linguistics* (pp. 1–9). CEUR-WS.

Bosco, C., Viviana, P., Bogetti, M., Conoscenti, M., Ruffo, G., Schifanella, R., et al. (2017). Tools and resources for detecting hate and prejudice against immigrants in social media. In *Proceedings of first symposium on social interactions in complex intelligent systems (SICIS), AISB convention 2017, AI and society*.

Brown, R. (2011). *Prejudice: Its social psychology*. Wiley.

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* (pp. 2833–2844). Association for Computational Linguistics.

Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., et al. (2020). *Detecting gender stereo types: Lexicon vs. Supervised learning methods* (pp. 1–11). Association for Computing Machinery.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In A. Armando, R. Baldoni, & R. Focardi (Eds.), *CEUR workshop proceedings*: *vol. 1816*, *Proceedings of the first Italian conference on cybersecurity* (pp. 86–95). CEUR-WS.

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

D'Errico, F., Paciello, M., & Amadei, M. (2018). Behind our words: Psychological paths underlying the un/supportive stance toward immigrants in social media. In *2018 IEEE 5th international conference on data science and advanced analytics* (pp. 649–656). IEEE.

D'Errico, F., Papapicco, C., & Taulë Delor, M. (2022). 'Immigrants, hell on board'. Stereotypes and Prejudice emerging from Racial Hoaxes through a Psycho-Linguistic Analysis. *Journal of Language and Discrimination*, *6*(2), 1–16.

D'Errico, F., & Poggi, I. (2012). Blame the opponent! Effects of multimodal discrediting moves in public debates. *Cognitive Computation*, *4*(4), 460–476.

D'Errico, F., Poggi, I., & Vincze, L. (2012). Discrediting signals. A model of social evaluation to study discrediting moves in political debates. *Journal on Multimodal User Interfaces*, *6*(3/4), 163–178.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.

van Dijk, T. A. (2016). Racism in the press. In N. Bonvillain (Ed.), *The Routledge handbook of linguistic anthropology* (pp. 384–392). New York: Routledge, Ch. 25.

Durrheim, K. (2012). Implicit prejudice in mind and interaction. In J. Dixon, & M. Levine (Eds.), *Beyond prejudice: Extending the social psychology of conflict, inequality and social change* (pp. 179–199). Cambridge University Press.

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!"?? analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Eds.), *CEUR workshop proceedings*: *vol. 2263*, *Proceedings of the sixth evaluation campaign of natural language processing and speech tools for Italian. Final workshop co-located with the fifth Italian conference on computational linguistics* (pp. 1–9). CEUR-WS.

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at IberEval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Eds.), *CEUR workshop proceedings*: *vol. 2150*, *Proceedings of the third workshop on evaluation of human language technologies for Iberian languages co-located with 34th conference of the Spanish society for natural language processing* (pp. 214–228). CEUR-WS.

Fields, C. (2016). *Stereotypes and stereotyping: Misperceptions, perspectives and role of social media*. Nova Science Pub. Inc.

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

Fiske, S. T., Cuddy, A. J., & Glick, P. (2006). Universal dimensions of social cognition: Warmth and competence. *TRENDS in Cognitive Sciences*, *11*(1), 77–83.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 85:1–85:30.

Francesconi, C., Bosco, C., Poletto, F., & Sanguinetti, M. (2019). Error analysis in a hate speech detection task: The case of HaSpeeDe-TW at EVALITA 2018. In *Proceedings of the sixth italian conference on computational linguistics (vol. 2481)* (pp. 1–7).

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

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., et al. (2019). SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 845–854). Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Kumar, R., Ojha, A. K., Lahiri, B., Zampieri, M., Malmasi, S., Murdock, V., & Kadar, D. (Eds.), (2020). *Proceedings of the second workshop on trolling, aggression and cyberbullying*. Marseille, France: European Language Resources Association (ELRA).

Kumar, R., Ojha, A. K., Zampieri, M., & Malmasi, S. (Eds.), (2018). *Proceedings of the first workshop on trolling, aggression and cyberbullying*. Association for Computational Linguistics.

Miceli, M., & Castelfranchi, C. (2000). The role of evaluation in cognition and social interaction. In *Human cognition and social agent technology* (p. 225). John Benjamins.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Poggi, I., & D'Errico, F. (2009). The mental ingredients of Bitterness. *Journal of Multimodal User Interface*, *3*(1), 79–86.

Poggi, I., D'Errico, F., & Vincze, L. (2011). Discrediting moves in political debate. In *Proceedings of the 2nd International Workshop on User Models for Motivational Systems: the affective and the rational routes to persuasion* (pp. 84–99).

Poletto, F., Basile, V., Bosco, C., Patti, V., & Stranisci, M. (2019). Annotating hate speech: Three schemes at comparison. In *Proceedings of the sixth italian conference on computational linguistics (vol. 2481)* (pp. 1–8). CEUR-WS.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., & Bosco, C. (2017). Hate speech annotation: Analysis of an italian twitter corpus. In *CEUR workshop proceedings*: *vol. 2006*, *Proceedings of the fourth Italian conference on computational linguistics* (pp. 1–6). CEUR-WS.

Polignano, M., Basile, V., Basile, P., de Gemmis, M., & Semeraro, G. (2019). AlBERTo: Modeling italian social media language with BERT. *Italian Journal of Computational Linguistics*, *5–2*.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897.

Ravasio, G., & Di Perna, L. (2020). GilBERTo: An Italian pretrained language model based on RoBERTa.

Risch, J., Stoll, A., Wilms, L., & Wiegand, M. (2021). Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments* (pp. 1–12). Duesseldorf, Germany: Association for Computational Linguistics.

Sánchez-Junquera, J., Chulvi, B., Rosso, P., & Ponzetto, S. P. (2021). How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, *11*(8), 3610.

Sanguinetti, M., Comandini, G., Nuovo, E. D., Frenda, S., Stranisci, M., Bosco, C., et al. (2020). HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. In V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), *CEUR workshop proceedings*: *vol. 2765*, *Proceedings of the seventh evaluation campaign of natural language processing and speech tools for Italian. Final workshop* (pp. 1–9). CEUR-WS.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An italian twitter corpus of hate speech against immigrants. In *Proceedings of the 11th conference on language resources and evaluation* (pp. 2798–2895). Miyazaki, Japan: ELRA.

Schmeisser-Nieto, W., Nofre, M., & Taulé, M. (2022). Criteria for the annotation of implicit stereotypes. In *Proceedings of the language resources and evaluation conference* (pp. 753–762). Marseille, France: European Language Resources Association.

Vaes, J., Latrofa, M., Suitner, C., & Arcuri, L. (2019). They are all armed and dangerous! *Journal of Media Psychology: Theories, Methods, and Applications*, *31*(1), 12–23.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (pp. 19–26). Stroudsburg, PA, USA: Association for Computational Linguistics.

Waseem, Z., Chung, W. H. K., Hovy, D., & Tetreault, J. (Eds.), (2017). *Proceedings of the first workshop on abusive language online*. Association for Computational Linguistics.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th conference on natural language processing* (pp. 1–10).

Ybarra, O., Burnstein, E., Winkielman, P., Keller, M., Manis, M., Chan, E., et al. (2008). Mental exercising through simple socializing: Social interaction promotes general cognitive functioning. *Personality and Social Psychology Bulletin*, *Feb*(34), 248–259.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 75–86).

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., et al. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020).