Natural Language Engineering (2022), 1–22
doi:10.1017/S1351324922000316

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Killing me softly: Creative and cognitive aspects of implicitness in abusive language online

Simona Frenda[1,2,*], Viviana Patti[1] and Paolo Rosso[2]

[1]Department of Computer Science, Università degli Studi di Torino, Turin, Italy and [2]PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain
*Corresponding author. E-mail: simona.frenda@unito.it

(Received 31 August 2021; revised 29 April 2022; accepted 19 May 2022)

Abstract

Abusive language is becoming a problematic issue for our society. The spread of messages that reinforce social and cultural intolerance could have dangerous effects in victims' life. State-of-the-art technologies are often effective on detecting explicit forms of abuse, leaving unidentified the utterances with very weak offensive language but a strong hurtful effect. Scholars have advanced theoretical and qualitative observations on specific indirect forms of abusive language that make it hard to be recognized automatically. In this work, we propose a battery of statistical and computational analyses able to support these considerations, with a focus on creative and cognitive aspects of the implicitness, in texts coming from different sources such as social media and news. We experiment with transformers, multi-task learning technique, and a set of linguistic features to reveal the elements involved in the implicit and explicit manifestations of abuses, providing a solid basis for computational applications.

Keywords: Abusive language detection; Figurative language; Hate speech; Stereotypes; Linguistic analysis

1. Introduction

Abusive language is a form of language that aims to create and enhance social tensions, provoking psychological and physical problems in victims, as well as dangerous forms of violence as offline attacks. For these reasons, in the last years, the need to detect the various forms of abusive language online (i.e., hate speech and stereotypes) is becoming an important topic for several public and private actors as well as activists.

However, the expression of abuses is not always explicit. Some messages, like Example (1),[a] implicitly contribute to mine the social tolerance towards the perceived *outgroup* (Fiske 1998).

(1) *I carabinieri hanno individuato come possibile spacciatore un 27enne del Marocco. La tipica #risorsa straniera, ammiro la madre! URL*
The Italian police have identified a 27-year-old from Morocco as a possible drug dealer. The typical foreign #resource, I admire the mother! URL

As suggested by Bianchi (2021), we can individuate two main dimensions of abusive language: an *evident* dimension that consists in the "verbal violence" that evokes the "physical violence"

Simona Frenda: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. Viviana Patti: Supervision, Project administration, Funding acquisition. Paolo Rosso: Supervision, Project administration, Funding acquisition.

[a]All the examples are extracted from the Italian hate speech corpus exploited in this work.

appearing explicitly aggressive and offensive; and a dimension *of propaganda* that aims at attesting the social identity presenting some roles or assumptions as normal and conventional, and appearing as a form of proselytism of negative idea. Thus, their consequences are various, both in victims and in society.

The exposure to harassment and microaggressions could provoke, in the long run, serious physical health issues such as cardiovascular diseases (Calvin *et al*. 2003) and immediate complex mental health issues such as depression and state of anxiety that might culminate in suicide (Nadal *et al*. 2014). It could also affect their willingness to engage in public and civic life, leaving, on the one hand, the communities which they are from unrepresented and depriving, on the other hand, the society where they live of a plurality of perspectives, useful in a democratic context.[b]

Moreover, although a causal link between the cyberharassment and hate crime is generally hard to demonstrate due to the difficulty to trace the particular texts that encourage the physical offenses, the risk of crime is assessed by victim surveys collected in the EU by the European Union Agency for Fundamental Rights,[c] by the systematic recording of crimes motivated by discriminatory bias, and by specific social studies that display the connection between hate speech spread online and crimes towards women, refugees, or religious and cultural minorities.

For instance, in the USA, Fulper *et al*. (2014) demonstrated the existence of a correlation between the number of rapes and the amount of misogynistic tweets per state. In London, linking police crime, census, and Twitter data, Williams *et al*. (2020) revealed a consistent association between online hate speech targeting race and religion and offline racially and religiously aggravated crimes. While in Germany, Müller and Schwarz (2021) demonstrated a connection among the political propaganda anti-refugees on Facebook, higher usage of social media, and crimes against refugees, as a clear effect of the echo chambers phenomenon.

The detection of hateful messages online, therefore, turns into a task of growing interest, and the techniques from Natural Language Processing (NLP) and Computational Linguistics (CL), could help to provide frameworks to formalize abusive language; unmask cognitive and linguistic processes implied in its comprehension; and propose models that allow machines to detect it automatically. However, the task of detecting abusive language is really challenging due to the different forms it can take. Waseem *et al*. (2017) emphasize the need to take into account two factors: the type of target (individuals or groups) and the degree of explicitness, which is established looking at the level of denotation and connotation of the hateful message.

Looking at the following examples:

(2) *user user . . . ma tutti i psicolabili stranieri li dobbiamo tenere noi in Italia?. . . se non sanno reprimere i loro istinti vanno tenuti segregati per non nuocere pi. . . invece stanno liberi. . .*
user user . . . but do we have to keep all the psychologically insane foreigners in Italy? . . . if they don't know how to repress their instincts they must be kept segregated in order not to harm anymore . . . instead they stay free . . .

(3) *Come osano chiedere il biglietto ai #migranti? Mandate subito i caschi blu dell'#ONU a vigilare sui diritti delle #risorse! Aggredito controllore sul #treno GTT URL*
How dare they ask #migrants for a ticket? Send the #ONU peacekeepers immediately to monitor the rights of #resources! Controller assaulted on #train GTT URL

we notice that in (2), the abusive language is unambiguous, whereas Example (3) involves sociocultural assumptions and ironic intention, that could be difficult to comprehend by humans with different backgrounds (Akhtar, Basile, and Patti 2020). Detecting correctly abusive language means, thus, understanding also the processes that make it indirect, such as typical cognitive bias and figurative language.

[b] https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf.
[c] https://fra.europa.eu/en/tools.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

Our main purpose in this paper is to shed light, by means of an ensemble of computational techniques, on the linguistic and cognitive elements involved in the implicit and explicit manifestations of abuse in texts, considering different media sources characterized by different ways of conveying abusive language.

In this regard, we focus on a benchmark dataset, called here HASPEEDE2020, released by the organizers of the HaSpeeDe shared task (Sanguinetti *et al*. 2020) at EVALITA 2020 for detecting *hate speech* and *stereotypes* in Italian. This shared task provides a suitable framework to investigate the different expressions of abusive language in a corpus of tweets and news headlines.

In HaSpeeDe, *stereotypes* are conceived as an orthogonal dimension of abusive language which do not necessarily coexist with hate speech: "*a standardized mental picture that is held in common by members of a group and that represents an oversimplified opinion, prejudiced attitude, or uncritical judgment.*" The proliferation of *oversimplified* and *uncritical* judgments about especially minorities causes the reinforcement of *outgroup homogeneity* perceived as different and, sometimes, in contrast with own in-group (Fiske 1998), like in:

(4) *#DecretoSalvini esatto e' buono anche per gli immigrati regolari che si vogliono integrare sul serio. la nostra cultura millenaria fara loro del bene.*
#DecretoSalvini exactly is even good for legal immigrants who want to integrate seriously. our millennial culture will do them good.

Concerning *hate speech*, the messages that aim at spreading or justifying hate, inciting violence, or threatening the freedom, dignity, and safety of individuals are considered hateful (Erjavec and Kovačič 2012; Sanguinetti *et al.* 2018). In HASPEEDE2020, the considered targets are the three most attacked minor communities in Italy: immigrants, Muslims, and Roma. Hate speech about these targets is often based on negative stereotyped ideas that categorize them as criminals or parasites. Such negative evaluations are very common also in traditional media such as news headlines where the hateful message is mostly implicit but still hurtful, for example in:

(5) *Il regno di immigrati e no global: "Ecco l'anticamera dell'inferno"*
The kingdom of immigrants and no global: "Here is the antechamber of hell"

The context of HaSpeeDe gives us the opportunity to investigate this social problem also from a multi-genres perspective, highlighting how hate speech and stereotypes are expressed in texts from different sources, and how, as orthogonal dimensions of abusive language, to interact between them.

As seen in Examples (1) and (3), hate speech could be made ambiguous by the use of figurative devices. Positive words ("resources," "admire," "peacekeepers," "rights") are sarcastically used to sugar-coat the real negative meaning and mask the stereotypes about immigrants. Some figures of speech, indeed, prove to be suitable for expressing hurtful opinions. For instance, in Frenda *et al*. (2022), sarcasm showed to be characterized by aggressive language, differently from other forms of irony that appear principally offensive in texts about immigration issues. Furthermore, Sanguinetti *et al*. (2018) noticed that, especially in case of negative and hateful opinions, social media users tend to be less explicit employing irony in their claims, in order to limit their exposure. Considering these previous studies, our main intuition is that the implicit abuses online can be manifested by the use of ironic language, and, thus, making aware the system of irony could improve the detection of abusive language.

HASPEEDE2020 contains texts coming from the dataset of tweets released during the first edition of HaSpeeDe in EVALITA 2018 (Bosco *et al*. 2018), and, since it was part of the Italian Hate Speech corpus (IHSC) described in Sanguinetti *et al*. (2018), they are already annotated as ironic and non-ironic. Therefore, we harmonized the annotation of HASPEEDE2020 labeling the missing instances, and, inspired by Cignarella *et al*. (2018), we provided also the annotation of sarcastic and non-sarcastic texts. Then, HASPEEDE2020 was extended with other tweets with the same

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

annotation coming from IHSC. This longer sample, called here HASPEEDE2020_EXT, was used for the statistical and experimental analyses.

Considering the availability of such a dataset, we wonder:

RQ1 What are the evident and hinted characteristics of hate speech and stereotypes in different textual genres?

RQ2 How do hate speech and stereotypes interact between them?

To this purpose, we carried various statistical and computational analyses exploiting the advantages of fine-tuning three Italian BERT-based Language Models (LMs) to detect hate speech and stereotypes in tweets and news headlines, putting emphasis on their "knowledge" derived from the data used for the training. Along with pre-trained LMs, we employed techniques of multi-task learning to make the classifier aware of other phenomena (i.e., stereotypes, hate speech, irony, and sarcasm). And, to show up the linguistic peculiarities of hate speech and stereotypes, we create a set of linguistic features aimed at capturing mainly connotative meanings, affective information, and syntactic patterns.

Answering these research questions leads us to examine deeply the creative and cognitive aspects of the abusive language online, providing a solid basis for the development of systems able to identify and prevent the spread of explicit and implicit manifestations of hate speech and stereotypes in texts coming from social media and news.

2. Implicit abuse and its open challenges

Although the attention on abusive language is recent in NLP and CL communities, the existing literature in this field is vast and not uniform. The subjective perception of the issue has caused various interpretations of the term *hate speech* and certain vagueness in the use of related terms such as abusive, toxic, dangerous, offensive, and aggressive language (Poletto *et al.* 2021; Vidgen and Derczynski 2021). Following the typology delineated by Waseem *et al.* (2017) and Poletto *et al.* (2021), we adopt the term *abusive language* as an umbrella term to enclose different expressions of abuse online.

Due to this lack of uniformity, most of the studies investigated on the detection of specific manifestations of abusive language, such as aggressiveness (Kumar *et al.* 2018; Carmona *et al.* 2018), flames (Lapidot-Lefler and Barak 2012), incivility (Rösner and Krämer 2016), cyberbullying (Dinakar, Reichart, and Lieberman 2011), offensiveness (Zampieri *et al.* 2019), toxicity (Taulé *et al.* 2021), misogyny (Fersini *et al.* 2018; Pamungkas *et al.* 2020) and racism (Waseem and Hovy 2016).

Recently, the focus of some benchmark competitions was extended to detection of various types of abuses in the same instances, providing, thus, new basis for more robust theoretical observations and computational models. Among them, Basile *et al.* (2019) organized the HatEval shared task at SemEval-2019 on hate speech and aggressive behavior detection in a multilingual corpus; the organizers of HASOC (Hate Speech and Offensive Content Identification in Indo-European Languages) at FIRE 2020 (Mandl *et al.* 2020) proposed to detect and distinguish offensive language, hate speech, and profanities in a multilingual dataset; Kumar *et al.* (2020) presented the second edition of TRAC (Trolling, Aggression and Cyberbullying) in 2020 on the detection of aggression and misogynistic aggression in multilingual data; Fersini, Nozza, and Rosso (2020) proposed a new edition of AMI (Automatic Misogyny Identification) at EVALITA 2020 asking participants to detect misogynistic texts and its aggressive attitude in Italian data, and finally, for the same occasion, Sanguinetti *et al.* (2020) presented the second edition of HaSpeeDe focused on detecting hate speech and stereotypes in Italian tweets and news headlines.

Nevertheless, the efforts towards combined analyses are still few. Clarke and Grieve (2017), for instance, investigated the functional linguistic variations between racist and sexist tweets of the corpus of Waseem and Hovy (2016), discovering that tweets against women tend to be more interactive and attitudinal than racist ones, addressed principally to persuade and argue

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

the discrimination reporting events. Lavergne *et al.* (2020) developed, for the second edition of HaSpeeDe, a competitive model based on multi-task learning approach to detect simultaneously hate speech and stereotypes, showing that injected knowledge about stereotypes improves the detection of hate speech only in tweets.

Moreover, the existing surveys on abusive language detection (Schmidt and Wiegand 2017; Fortuna and Nunes 2018) underline the necessity to computationally approach the implicitness of toxic discourses, especially in the cases where these are disguised by sarcasm, euphemism, rhetorical questions, litotes, or where there are no explicit accusations, negative evaluations, or insults. This kind of implicitness eludes the offensiveness of the text, making its recognition hard, especially for machines (Nobata *et al.* 2016; Frenda 2018; MacAvaney *et al.* 2019).

Additionally, the evaluation process based on specific test sets and measures, such as accuracy and F1-score, could overestimate the model performance without revealing particular weak points. To this purpose, Röttger *et al.* (2021) developed a suite of functional tests, called HATECHECK, to evaluate deeply the models for abusive language detection in English. In particular, on the basis of previous works and interviews, they elaborated 29 functional tests that cover the most common challenges in hate speech detection, such as negation, slurs, pronoun reference, threatening language, counter speech, and spelling errors. The usefulness of this tool was confirmed by Vidgen *et al.* (2021), who tested their dynamic approach of dataset generation, proving the robustness of its model trained in various rounds. A dynamic approach, indeed, allows coping with problems when the work is conducted, discussing with expert annotators, extending and ameliorating step by step the training set, annotated taking into account different types and targets of hate.

The availability of various vulnerable targets helps the classifier to generalize better the presence of hate, without excluding the identification of abusive language towards unseen groups. An example comes from Talat, Thorne, and Bingel (2018), where authors proposed a multi-task learning based model with the aim to bridge differences in annotation and data collection such as different annotation schemes, labels, or geographic and cultural influences from data sampling.

As suggested by Jurgens, Hemphill, and Chandrasekharan (2019), NLP community needs, indeed, to expand its efforts to recognize infrequent abuses (taking into account especially the context where these abuses occur) and detect subtle abuses that could be manifested as benevolent stereotyping, condescension, minimization, or disparity in treatment of social groups. In addition, Wiegand, Ruppenhofer, and Eder (2021) identified specific subtypes of implicit abuse analyzing various benchmark datasets in English: stereotypes, perpetrators, comparisons, dehumanization, euphemistic constructions, call-for-action, multimodal abuse, and all the phenomena that require world knowledge and inferences such as jokes, sarcasm and rhetorical questions.

Some of these subtypes have been identified by scholars as problematic challenges in abusive language detection, demonstrating that only its explicit manifestations are *understood* by current classifiers (supervised and unsupervised). For instance, Van Aken et al. (2018) proposed a detailed error analysis of an ensemble classifier's performance in a Wikipedia[d] and Twitter (Davidson *et al.* 2017) dataset, individuating specific phenomena that make abusive language difficult to recognize: lack of explicit offenses (such as swearwords), idiosyncratic expressions, rhetorical questions, metaphorical, and ironic language. As shown also by Wiegand, Ruppenhofer, and Kleinbauer (2019), the performance of classifiers in presence of implicit abuse decreases considerably, with some exception regarding those cases where the sampling process introduces data bias in the training and test set. These analyses that take into account the explicit and implicit portion of abusive documents are carried out looking at the vocabulary of the corpora: a document contains explicit abusive language if it includes at least one word from a lexicon of abusive words (Wiegand *et al.* 2018). The same approach is employed to OLID/OffensEval dataset (Zampieri *et al.* 2019) by Caselli *et al.* (2020) as a basic analysis to reflect about the notions of explicit/implicit and

[d]This dataset has been published by Google Jigsaw in December 2017 in the context of Toxic Comment Classification Challenge on Kaggle.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

offensive/abusive and then propose a new annotation on OLID/OffensEval creating AbuseEval v1.0. As expected, the authors showed that the documents annotated as offensive in OffensEval overlaps largely with the documents annotated as explicitly abusive in AbuseEval and that the identification of the implicit abuse is more difficult than the explicit one.

Coping with implicit phenomena is necessary to make systems able to understand these messages that have a strong abusive effect but very weak offensive forms. Bowes and Katz (2011), for example, noted that the victims of sarcastic utterances do not perceive the expression as humorous, differently from the aggressors' point of view, and not less polite than the literal counterpart. This study contradicts the line of some scholars that stress the hypothesis that considers ironic language as a device to mute the negative meaning (Dews and Winner 1995). In this regard, Pexman and Olineck (2002) proposed a pragmatic analysis of *ironic insult* and *ironic compliment*: the former is perceived as more polite whereas the latter as mocking and sarcastic. Speakers, in fact, tend to criticize someone lowering the social cost of doing so, and ironic language seems appropriate to conceal the abuse.

In spite of the theoretical literature clearly describes the implicitness of abusive language, the computational efforts that could support it are few. To our knowledge, only stereotypes and metaphors have been exploited for abusive language detection. For instance, Lemmens, Markov, and Daelemans (2021) proved the contribution of hateful metaphors as features for the identification of the type and target of hate speech in Dutch Facebook comments using models based on classical machine learning and transformers.

In this context, our contribution aims to bring to light, by means of statistical and computational analyses, the *invisible* processes that characterize indirect abusive language in terms of cognitive bias and ironic language.

3. Figurative and cognitive aspects: Statistical analysis of the corpus

To answer our research questions, the HaSpeeDe context seems to provide a suitable framework for the Italian language. The HaSpeeDe shared task proposed by Sanguinetti *et al.* (2020) consists of three sub-tasks[e]:

- Task A (*Hate Speech Detection*) is a binary classification task that, like in the first edition in 2018, asks participating systems to predict whether a text contains hate speech (*hs* or *non-hs*) towards a given target (immigrants, Muslims and Roma);
- Task B (*Stereotype Detection*) is a binary classification task aimed at determining the presence or the absence of a stereotype (*stereo* or *non-stereo*) towards the same targets;
- Task C (*Identification of Nominal Utterances*) is a sequence labeling task to recognize NUs in texts previously predicted as hateful.

Considering this context of analysis on the implicitness of abusive language, in this work we face computationally only the first two tasks, exploiting the annotation provided for Task C for additional analysis.

3.1 Description of the dataset

The HaSpeeDe2020 dataset contains tweets and news headlines: a part gathered from other existing corpora and another collected in the last years from social media and newspapers online. All these data were annotated using the guidelines defined for IHSC.[f]

Considering the aim of the organizers of HaSpeeDe of encouraging the development of more robust systems of detection in cross-genre contexts, they proposed two test sets in the competition:

[e]Refer to the overview of the organizers Sanguinetti *et al.* (2020) for more details.
[f]https://github.com/msang/hate-speech-corpus/blob/master/GUIDELINES.pdf.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

Table 1. Distribution of Labels in HASPEEDE20_EXT

Set	hs	non-hs	stereo	non-stereo	iro	non-iro	sarc	iro non-sarc	Total
Tweets									
TRAIN_TW	2,766	4,073	3,042	3,797	–	–	–	–	6,839
TRAIN_TW_EXT	3,035	5,226	3,554	4,707	1,806	6,455	1,111	695	8,261
TEST_TW	622	641	569	694	361	902	239	122	1,263
#token/text	31.74	25.55	30.31	26.1	30.16	27.27	30.39	27.52	–
News									
TEST_NW	181	319	175	325	40	460	21	19	500
#token/text	15.43	16.43	15.73	16.25	16.9	16.0	17.95	15.99	–

one composed of tweets (TEST_TW) and another composed of news headlines (TEST_TW), whereas the training set (TRAIN_TW) of HASPEEDE2020 consists only of tweets.

About the composition of TRAIN_TW and TEST_TW, a part of tweets comes from the Twitter dataset released in the first edition of HaSpeeDe in 2018 (and partially derived from IHSC); the rest instead has been gathered for the Italian hate speech monitoring project "Contro l'Odio" (Capozzi *et al*. 2019). In particular, only data posted between September and December 2018 were included in TRAIN_TW, whereas TEST_TW contains the tweets posted between January and May 2019.

The news headlines about immigrants related events were used only in the second test set (TEST_NW). These data were retrieved between October 2017 and February 2018 from online newspapers such as *La Stampa*, *La Repubblica*, *Il Giornale*, *Liberoquotidiano*.

Taking into account this composition of HASPEEDE2020 and the fact that some tweets are already annotated as ironic and non-ironic (*iro* and *non-iro*), we harmonized the annotation of HASPEEDE2020 labeling the missing instances, and providing also the annotation of sarcastic and non-sarcastic (*sarc* and *non-sarc*) texts. For this last label, we followed the schema of annotation used for creating the dataset released for the IronITA shared task organized by Cignarella *et al*. (2018) at EVALITA 2018.

In IronITA, irony is defined as a figurative language device that conveys a meaning that is secondary or opposite to the literal one. Linguistic literature places irony among *metalogic figures* that are the figures that modify the logic value of the utterance, breaking the maxim of quality (Grice 1975) and affecting the literal meaning (Garavelli 1997). Instead, sarcasm is conceived as a type of irony that aims to mock and scorn a victim, and that, differently from other forms of irony, is used to ridicule a specific target (Lee and Katz 1998).

Finally, TRAIN_TW was extended with other instances with the same annotation from IHSC, to have a longer sample for the analyses. This extended version is called here HASPEEDE2020_EXT.

Table 1 shows the data distribution in HASPEEDE2020_EXT (as well as the increment of tweets from TRAIN_TW to TRAIN_TW_EXT) and the average of number of tokens per instance in each class (`#token/text`). Taking into account the different textual genres of our dataset, we calculated the average separately for tweets and news headlines. Table 2 reports some examples from TRAIN_TW_EXT.

3.2 Statistical analysis

Exploiting the various labels for each instance, we applied a statistical analysis to study the association between ironic language (irony and sarcasm) and abusive language (hate speech and stereotype) interpreted as nominal variables of a population. In particular, we computed:

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

Table 2. Examples from HASPEEDE20_EXT

hs	stereo	iro	sarc	text
0	0	0	0	*Iniziano ritorsioni contro Governo: sindaco PD espropria case a italiani per darle ai migranti URL* →Retaliation against the government begins: PD mayor expropriates homes from Italians to give them to migrants URL.
0	1	1	0	*Fanno degli scambi con i documenti. . .#quartogrado tra Ucraini e Rumeni.* →They do exchanges with documents . . . #fourthgrade between Ukrainians and Romanians.
1	1	1	1	*Parlamentari #sinistri dicono che i migranti arrivano con lievi patologie e s'ammalano gravemente in Italia: benvenuto #colera ! URL* →#Left/Sinister parliamentarians say that migrants arrive with minor illnesses and become seriously ill in Italy: welcome #colera ! URL

Table 3. *p*-values/Yule's Q values between Ironic and Abusive Language

	Tweets		News	
	hs	stereo	hs	stereo
irony	0.00/-0.12	0.00/0.10	**0.00/0.72**	0.00/0.70
sarcasm	**0.00/0.24**	0.00/0.15	0.85/0.07	0.58/0.19
non-sarcastic irony	0.00/-0.27	0.75/0.01	**0.00/0.70**	0.00/0.65

- the χ^2 test of independence that, by means of the interpretation of *p*-value, gives information on the existence or not of significant relations between nominal variables;
- the Yule's Q to indicate if the association between two binary variables is positive (values close to 1), negative (values close to -1), or null (values close to 0).

To reject the null hypothesis (hypothesis that the variables are independent) of the χ^2 test of independence, the *p*-value should be minor than the significance level set by convention to 0.05, and to calculate the *p*-value, we considered a degree of freedom based on the number of observations. The results of this analysis are reported in Table 3.

Table 3 shows that in tweets, the association between sarcasm and abusive language reports high scores, especially in the case of hate speech. Differently from sarcasm, the values related to the relation between irony and stereotypes are lower and in the case of non-sarcastic irony the relation is even absent.

About the genre of news headlines, the association between hateful and ironic language appears stronger than in tweets, especially in the cases where irony is not sarcastic. Although the values related, particularly, to news headlines are based on very few data (Table 1), this analysis gives us a first look of the possible characteristics of indirect language in messages containing hate speech and stereotypes in different textual genres.

We propose the same analysis between the two dimensions of abusive language analyzed in this work.

Table 4 shows a strong association between hate speech and stereotypes, confirming the fact that, especially in implicit contexts such as news headlines, abusive language is characterized mainly by negative stereotypes that support the intolerance towards specific groups (see Table 2).

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

Table 4. *p*-values/Yule's Q values between Hate Speech and Stereotypes

	Tweets	News
	`stereo`	`stereo`
`hs`	0.00/**0.80**	0.00/**0.96**

4. Computational analyses

The statistical analysis, that partially answer our research questions, confirms our initial intuition on the possible implicit manifestations of abusive language through sarcasm in informal texts and through stereotyped ideas in formal contexts such as news headlines. In order to bring to light the implicit and explicit characteristics of hate speech and stereotypes in different textual genres [RQ1], and understand well how these two dimensions of hate interact between them [RQ2], we carried out a battery of computational experiments on abusive language detection exploiting the HaSpeeDe framework.

We used the same test sets (TEST_TW and TEST_NW) and evaluation measure proposed at the shared task, to compare the results with attested baselines and other models. However, for the training phase, we exploit the extended version of the training set: TRAIN_TW_EXT.

In particular, we designed three different set of systems based on:

1. simple fine-tuning three LMs for Italian (FT_MODEL),
2. combination of the LMs' knowledge with the awareness derived from the simultaneous learning of related tasks (MTL_MODEL),
3. combination of the knowledge derived from MTL_MODEL with specific linguistic features (MTL_MODEL+FEATURES).

FT_MODEL. Considering the popularity in recent years of transformers, we selected three LMs trained on different genres of texts to reveal the contribution of transfer learning in a cross-genre context for abusive language detection.

a. AlBERTo (Polignano *et al*. 2019) is trained on TWITA (Basile, Lai, and Sanguinetti 2018), a large dataset collecting Italian tweets from February 2012.[g]
b. Italian BERT (ItBERT)[h] is trained on Wikipedia dump and various texts from the OPUS corpora (Tiedemann and Nygaard 2003) for a total size of 13gb.
c. Italian BERT XXL (ItBERTXXL)[i] is a sort of extended version of ItBERT. It is trained on the same data from OPUS of ItBERT and on additional documents coming from the Italian part of the OSCAR corpus (Ortiz Suárez, Sagot, and Romary 2019) for a final size of 81gb.

MTL_MODEL. As shown in Tables 3 and 4, abusive language tends to be expressed with creative or cognitive aspects in respect to the formal or informal context. Therefore, in this set of experiments we aim:

(1) to quantify the impact of additional knowledge related to ironic language in abusive language detection,

[g]The model used in this work is trained on 200M tweets published from 2012 to 2015: https://github.com/marcopoli/AlBERTo-it.
[h]https://huggingface.co/dbmdz/bert-base-italian-uncased.
[i]https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

(2) to comprehend deeply the interaction between hate speech and stereotypes, even in a neural network context.

To this purpose, we employed an approach based on multi-task learning. At the computational level, the advantages derived from the use of multi-task learning techniques, such as the *hard parameter sharing*, are various. Firstly, this technique gives systems more evidences to evaluate if a feature is relevant or not, focusing strictly on the most relevant ones for each task. Then, the hard parameter sharing allows a better generalization for each task: learning simultaneously more tasks means to find a representation that is appropriate for learning all the tasks, reducing consequently the over-fitting on the original task (Baxter 1997).

MTL_MODEL+FEATURES. In the last set of experiments, we aim to examine the relevance of specific linguistic features in texts containing hate speech and stereotypes, and to estimate their contribution ahead of a classification task.

To this purpose, we designed:

- a set of linguistic features to capture information related to style of writing, syntax, lexical semantics, and pragmatics such as emotions and sentiment,
- a neural network that converges in a unique model the knowledge coming from the LMs in the MTL context, and the specific knowledge derived from dedicated linguistic features.

4.1 Neural network architecture

In the first set of experiments (FT_MODEL), we simply fine-tuned the LMs on hate speech and stereotypes classification tasks, taking into account the pooled output of the BERT-based model, adding a dropout layer to prevent the over-fitting, a dense layer with standard ReLU activation, and a final dense layer to get the class-related probability employing a Sigmoid function. As optimizer, we used Adam with a really low learning rate (0.00001) found by means of a specific callback function.[j] Finally, to minimize the loss function during the training, we used the binary cross-entropy function for binary classification provided by *keras* library.

The same structure and hyperparameters are applied for the second set of experiments (MTL_MODEL). For the MTL context, on the top of the previous network, we added two final dense layers that employ Sigmoid function to get one output for each task. In accordance with the standard BERT input representation (Devlin *et al.* 2018), the text is represented in both networks as tokens, segments, and masked input. To load the pre-trained models for *TensorFlow* environment and tokenize the texts for creating tokens-input, we used *transformers* library[k].

In the last set of experiments (MTL_MODEL+FEATURES), we employed the same network of MTL_MODEL, adding a concatenate-layer that combines the pooled output of BERT-based models with a features' vector representation input-layer. Before the concatenation, we applied the batch normalization technique to the features' input-layer to standardize the layer and stabilize the learning process. For the MTL context, also here we have two output layers, one for each task.

To weight the majority of extracted features, we used the TF-IDF (term frequency-inverse document frequency) measure calculated for each word of the vocabulary (TF-IDF_DICT). Other semantic and pragmatic features are weighted considering the scores of polarity and the cosine similarity. These weights have been standardized using *MinMaxScaler* of *scikit-learn* with default range of scaling.

To create the TF-IDF_DICT and the word embedding model used to compute the cosine similarity, we preprocessed the texts: deleting URLs and symbols like @ and # to maintain the lexical information of hashtags and usernames; tokenizing and lemmatizing words using

[j] http://di.unito.it/lrfinder.
[k] https://huggingface.co/transformers/.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

the implementation of TreeTagger[l] for python[m]; and removing stopwords[n] to retain lexical significant words.

4.2 Linguistic features

To create the set of features, we took inspiration from previous works that define specific patterns to identify irony, sarcasm (Hernández Farías, Patti, and Rosso 2016; Cignarella *et al.* 2020), and abusive language (Frenda *et al.* 2020).

Style related features. This set contains punctuation marks and patterns of negation. Punctuation (`punct`) is commonly used to express the intended meaning of the message. For instance, users use quotation marks to point out the opposite of the literal meaning of a word, such as *"risorsa"* ("resource"). Negation (`negation`) proved to play an important role in the process of comprehension of figurative language (Giora, Givoni, and Fein 2015; Karoui *et al.* 2017). In the features' vector, they are represented by the sum of their TF-IDF values in the text.

Syntax related features. This set involves specific syntactic dependencies expressing adverbial locutions (`adv_loc`), intensifiers (`intens`), discourse connections (`disc_conn`), mentions (`mention`), and nominal phrases (and the number of nominal phrases in the tweet) (`nom_phrase` and `num_nom_phrase`). To extract these features, we used *spacy-udpipe* library with TWITTIRò model specified for short texts in Italian[o] (Cignarella *et al.* 2019), and to retrieve their weights, we exploited TF-IDF_DICT.

Lexical semantics-related features. This set is composed of:

- Lexical information about offensive language extracted exploiting the HurtLex[p] multilingual lexicon (Bassignana, Basile, and Patti 2018). HurtLex was created from the Italian lexicon "Le Parole per Ferire" by Tullio de Mauro, and the words in the lexicon are classified in 17 types of offenses (see Table 5) enclosed in two macro-categories: *conservative*, that are the words with literally offensive sense; and *inclusive*, that are the words with not literally offensive sense but that could be used with negative connotation. To extract these features, we used the featurizer[q] created specifically for this lexicon that put the attention on the categories. Therefore, to represent them in the features' vector, we computed the sum of the TF-IDF values of all words in the text belonging to each category.
- Semantic information about incongruities and similarities revealed by words and pairs of words in the text. These features are extracted considering the variability of the TF-IDF weights of the words in the text by means of the standard deviation (σ) and the coefficient of variation (`cv`), the average of weights (`avg`), and the maximum (`max`), minimum (`min`), and median (`med`) values of the list of the TF-IDF values of words (`W`) and bigrams of words (`B`) in a text. The values related to bigrams are computed using the weights' normalization on the scores of maximum and minimum (`C1`) and of standard deviation and average (`C2`). Moreover, we calculated the similarity (cos (θ)): 1 between pairs of words (vector of bigram of words) and the sentence context (corresponding to sentence vector) (cos (θ)_`BS`), and 2 between the bigrams of words within the sentence (cos (θ)_`BB`). To represent them in the features' vector, we computed σ, the coefficient of variation, the average, and maximum, minimum, and median scores of lists of cosine similarity values. The word embedding

[l] Using this tool the numbers are replaced by `@card@` tag.
[m] https://treetaggerwrapper.readthedocs.io/en/latest/.
[n] The used list of stopwords is available in http://di.unito.it/stopwordsit.
[o] http://di.unito.it/twittirotreebank.
[p] http://hatespeech.di.unito.it/resources.html.
[q] https://github.com/valeriobasile/hurtlex.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

Table 5. HurtLex Categories

Category	Description
PS	Ethnic Slurs
RCI	Location and Demonyms
PA	Profession and Occupation
DDP	Physical Disabilities and Diversity
DDF	Cognitive Disabilities and Diversity
DMC	Moral Behavior and Defect
IS	Words Related to Social and Economic advantages
OR	Words Related to Plants
AN	Words Related to Animals
ASM	Words Related to Male Genitalia
ASF	Words Related to Female Genitalia
PR	Words Related to Prostitution
OM	Words Related to Homosexuality
QAS	Descriptive Words with Potential Negative Connotations
CDS	Derogatory Words
RE	Felonies and Words Related to Crime and Immoral Behavior
SVP	Words Related to the Seven Deadly Sins of the Christian Tradition

model, created for computing these values has been created following the methodology presented in Frenda *et al.* (2022).

Pragmatics related features. This set consists of affective information extracted exploiting Sentix and Emolex dictionaries.

- To extract the sentiment of the text, we used Sentix[r] (Basile and Nissim 2013), that contains for each lemma scores about positive and negative sentiment, polarity, and intensity. Using this information, we calculated the average of positive and negative score of words in the text (`avg_positive` and `avg_negative`), σ of polarity, and the intensity average (`avg_intensity`).
- To capture emotions and feelings, we used EmoLex[s] (Mohammad and Turney 2013), a multilingual lexicon containing for each lemma information about the 8 principal emotions of Plutchik (Plutchik and Kellerman 2013). Inspired by Plutchik (2001), we exploited the wheel of emotions to capture in the message the *principal emotions* (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), the primary dyads or *feelings* (aggressiveness, optimism, love, submission, awe, disapproval, remorse, contempt), and the variability of opposite emotions/feelings by means of σ. Emotions and feelings are represented

[r] http://valeriobasile.github.io/twita/sentix.html.
[s] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

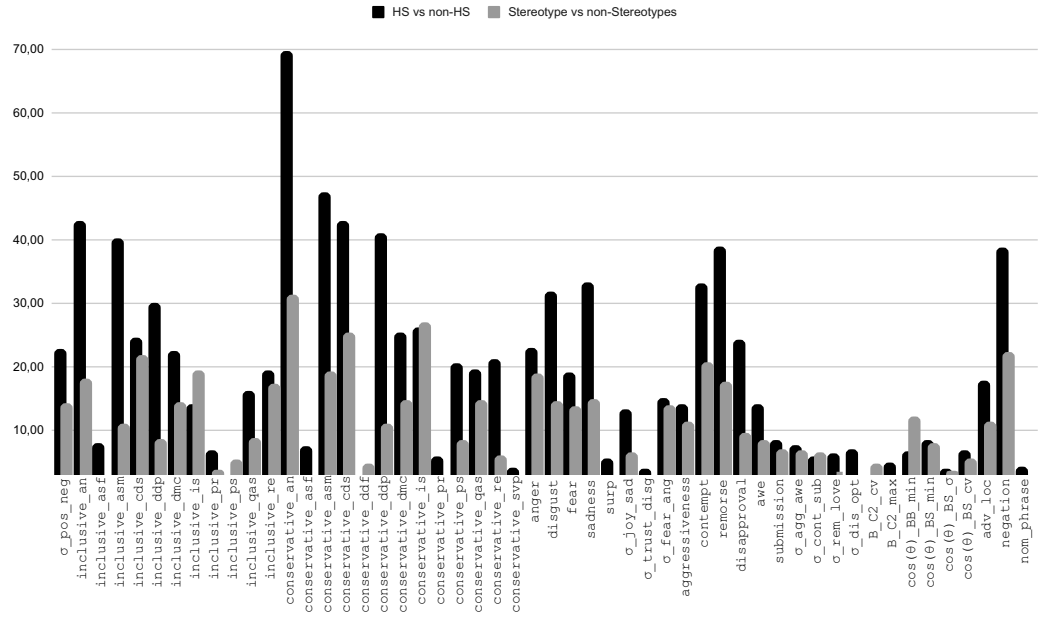

Figure 1. Relevant Characteristics in TRAIN_TW_EXT.

in the features' vector by the sum of the TF-IDF values of the words related to each emotion/feeling in the text.

4.2.1 Features relevance

Figure 1 shows the relevance of the sets of features analyzed for each phenomenon, computed by means of χ^2 values. In particular, we report the most relevant features with a χ^2 greater than 3.

Looking at Figure 1, we can notice that Hate Speech and Stereotypes are characterized by very similar features. In general, both are featured by negative emotions and feelings (anger, awe, disgust, aggressiveness, fear) and by offensive words with conservative and inclusive interpretation. Some categories of offenses related especially to animals, physical disabilities or diversity, behaviors/morality, and general swear words are more relevant in hate speech, whereas in stereotyped messages the offenses more significant are linked in particular to economic and social issues, cognitive, and ethnic sphere, even if in a more indirect way.

At semantic level, the minimum score of similarity between the bigrams of words within the sentence (cos (θ)_BB) appears to be relevant in stereotypes recognition. This specific feature brings out the semantic incongruity in the text: a common technique used to express irony (Riloff *et al*. 2013; Joshi, Sharma, and Bhattacharyya 2015; Pan *et al*. 2020).

Finally, both phenomena appear characterized by specific syntactic patterns, such as negation and adverbial locutions. From a manual examination, we found that the former are used especially to mark some characteristics of outgroup, juxtaposing it sometimes with the in-group, while the latter tend to increase the intensity of some beliefs or make the sentences mainly nominal. Here are some examples of the presence of these markers:

(6) *Ora ankio andrò ad emigrare dato che qui sarà tutto occupato da stranieri. tanto di noi italiani non gliene frega nessuno. . .. vergognaaa*
Now me too I will go to emigrate since everything here will be occupied by foreigners. *so much of* us Italians do *not* give a damn. . .. shame on

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

(7) *I nostri migranti non erano assolutamente come questa gentaglia qui! Continuare a fare questo paragone un'offesa per tutti quelli che si sono rotti mani e schiena nelle miniere e nelle fabbriche e vivevano nascosti il resto del tempo.*
Our migrants were absolutely *not* like this scum here! Continuing to make this comparison is an offense to all those who broke their hands and backs in mines and factories and lived in hiding the rest of the time.

4.3 Experimental setting

Our computational analyses, interpreted in this work as classification tasks of the texts containing hate speech (Task A) and stereotypes (Task B), are performed using the same setting and hyperparameters for all the sets of experiments.

To evaluate the performance of the models during the training, we used 20% of TRAIN_TW_EXT as validation set. The systems have been trained with a maximum of 20 epochs for each run and a batch size of 32 for each epoch. To avoid problems of over-fitting, we used the early stopping function, monitoring the loss value obtained on the validation set, and to obtain reproducible results, we set a seed function.

In regard to MTL_MODEL+FEATURES, we evaluated the performance using both the designed features (FEAT) and selected features (SELECTEDFEAT) for each detection task. To select the best features, we considered their χ^2 value (greater than 10) for a total of 27 features for hate speech and 25 for stereotype detection.

Exploiting the HaSpeeDe framework, each model, for both Task A and Task B, has been evaluated using the test sets (TEST_TW and TEST_NW) and evaluation measure proposed by the organizers of the shared task: F1-macro as average score of F1 value of each class. Moreover, to assess the characteristics of hate speech and stereotypes in tweets and news headlines, we compared the obtained results with the straightforward baselines provided by the organizers and the best scores obtained by the teams first ranked in the competition:

- Baseline_MFC (Most Frequent Class) that assigns to each instance the majority class of the respective task.
- Baseline_SVC that classifies the texts using an SVM algorithm with unigrams and chargrams (2-5) weighted with TF-IDF.
- TheNorth team (Lavergne *et al.* 2020) used a simple neural network that fine-tunes UmBERTo[t] LM using specifically a linear layer with a softmax on top of the CLS token, and applying a novel technique of layer-wise learning rate. TheNorth submitted two runs for each task. The first obtained simply fine-tuning UmBERTo and the second using a multitasking approach that exploits the possible correlation between texts containing hate speech and texts expressing stereotypes about the targets.
- CHILab team (Gambino and Pirrone 2020) experimented transformer encoders in the first run, creating specifically two transformer/convolution blocks for each input (texts and Part-of-Speech or PoS tags) averaged through max pooling and processed finally by a dropout and dense layer to obtain the predictions, and a depth-wise Separable Convolution techniques in the second one. CHILab also used additional tweets taken from TWITA by means of some keywords extracted from the provided training set to extend the embedding layer of their model.

[t] https://github.com/musixmatchresearch/umberto.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

4.4 Results and discussion

Observing the results in Table 6[u] obtained within the first set of experiments, we can notice interesting correlations in both tasks: the systems using the ItBERTXXL model perform well in tweets (0.788 and 0.764), whereas the systems using ItBERT in news headlines (0.674 and 0.733). AlBERTo, differently from what we expected, contribute to improve the detection of stereotypes in the news genre when the classifier is aware of irony (0.752). Evidently, the generalization about the style of short texts coming from AlBERTo and the knowledge derived from the training on ironic texts help the system to catch some indirect dependencies that make it able to recognize the stereotypes in a very difficult context such as news headlines.

In order to quantify the impact of linguistic and cognitive aspects, we computed the percentage of Δ between the values of F1-score of best models (underlined in Table 6) and the best baseline (*Baseline_SVC*). For a better visualization, we report these values in Table 7.

About the expression of stereotypes, differently from the statistical results reported in Table 3, the percentage of variation from baseline model (12.41%) suggests that it is characterized by specific patterns typical also of ironic language, such as the reference to secondary or indirect meanings, principally in less spontaneous texts like news headlines. The opposite trend is visible in hate speech detection, that shows higher values of Δ when the classifier is aware of sarcasm in both genres, confirming the previous statistics about the use of a sharper form of irony especially in tweets [RQ1].

Looking at the contribution of the mutual information between hate speech and stereotypes, we can notice that, actually, only the detection of hate speech takes advantage of knowledge derived from the simultaneous learning of stereotype detection. And, as we can see in Table 7, the values of Δ in Task A (12.07% and 11.59%) are very high regardless the textual genre. During the HaSpeeDe shared task, only TheNorth team experimented an approach multi-task learning, showing that the knowledge about stereotypes could improve the identification of hate speech in tweets ($\Delta = 12.20\%$).

These results allow us to interpret better the strong association emerged in Table 4, proving that even if hate speech could be expressed using negative stereotypes to reinforce or justify the message, the same is not true in reverse [RQ2].

Finally, observing in Table 6 the role of features in the different experiments, we notice that specific linguistic information contributes to increase above all the performance of ItBERT-based models in news headlines. News headlines, in general, are characterized by a different style of writing, less spontaneous than texts coming from social media. To examine their characteristics, we exploited the coarse-grained annotation of NUs provided for Task C, extracting from them the most frequent trigrams of words: "*basta balle ecco*" ("no more lies here"), "*via i migranti*" ("out the migrants"), and "*immigrati la verità*" ("immigrants the truth"). These are involved in specific syntactic contexts, like:

(8) *Immigrati, ammiraglio brutale: ora basta balle. "Ecco chi trama contro l'Italia, serve una guerra"*
Immigrants, brutal admiral: no more lies now. "Here is who is plotting against Italy, we need a war"

(9) *C'è la scuola, via i migranti: "Siamo contrari allapartheid ma ora serve più sicurezza"*
There is the school, out the migrants: "We are against apartheid but now we need more security"

The identified NUs remember a peculiar political rhetoric that aims to feed the intolerance against immigrants, called specifically *Slogan-like NUs* by Comandini and Patti (2019). The style of news headlines, thus, makes the detection especially of hate speech harder, and although the

[u] Taking into account the scope of this work, we reported in Table 6 only the best scores obtained by TheNorth and CHILab.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

Table 6. Results for Task A and B in Tweets and News Headlines

		Task A		Task B	
System	Id	F1_Tw	F1_Nw	F1_Tw	F1_Nw
Baseline_MFC		*0.337*	*0.389*	*0.355*	*0.394*
Baseline_SVC		*0.721*	*0.621*	*0.715*	*0.669*
TheNorth	2	**0.809**	–	0.768	–
TheNorth	1	0.790	–	**0.772**	–
CHILab	1	0.789	**0.774**	0.761	0.720
FT_AlBERTo		0.753	0.624	0.744	0.682
FT_ItBERT		0.766	0.674	0.725	0.733
FT_ItBERTXXL		0.788	0.540	0.764	0.671
MTL_AlBERTo					
(HS/Stereotype-Irony)		0.748	0.600	0.713	**0.752**
(HS/Stereotype-Irony)+Feat		0.744	0.621	0.724	0.727
(HS/Stereotype-Irony)+SelectedFeat		0.770	0.581	0.718	0.710
(HS/Stereotype-Sarcasm)		0.780	0.616	0.723	0.715
(HS/Stereotype-Sarcasm)+Feat		0.748	0.601	0.711	0.697
(HS/Stereotype-Sarcasm)+SelectedFeat		0.773	0.568	0.732	0.699
(HS-Stereotype)		0.770	0.676	0.703	0.719
(HS-Stereotype)+Feat		0.762	0.632	0.733	0.683
(HS-Stereotype)+SelectedFeat		0.772	0.635	0.724	0.699
MTL_ItBERT					
(HS/Stereotype-Irony)		0.781	0.585	0.762	0.687
(HS/Stereotype-Irony)+Feat		0.768	0.557	0.743	0.743
(HS/Stereotype-Irony)+SelectedFeat		0.761	0.623	0.751	0.718
(HS/Stereotype-Sarcasm)		0.746	0.571	0.724	0.697
(HS/Stereotype-Sarcasm)+Feat		0.760	0.659	0.720	0.713
(HS/Stereotype-Sarcasm)+SelectedFeat		0.770	0.619	0.724	0.714
(HS-Stereotype)		0.761	0.610	0.731	0.692
(HS-Stereotype)+Feat		0.808	0.572	0.729	0.692
(HS-Stereotype)+SelectedFeat		0.768	0.633	0.760	0.706
MTL_ItBERTXXL					
(HS/Stereotype-Irony)		0.778	0.598	0.767	0.747
(HS/Stereotype-Irony)+Feat		0.756	0.671	0.744	0.719

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

Table 6. Continued

		Task A		Task B	
System	Id	F1_Tw	F1_Nw	F1_Tw	F1_Nw
(HS/Stereotype-Irony)+SelectedFeat		0.758	0.647	0.755	0.746
(HS/Stereotype-Sarcasm)		0.789	0.676	0.733	0.688
(HS/Stereotype-Sarcasm)+Feat		0.782	0.621	0.749	0.710
(HS/Stereotype-Sarcasm)+SelectedFeat		0.782	0.618	0.742	0.696
(HS-Stereotype)		0.792	0.615	0.732	0.727
(HS-Stereotype)+Feat		0.796	0.630	0.769	0.712
(HS-Stereotype)+SelectedFeat		0.791	0.693	0.770	0.715

Table 7. Percentages of Variation

		Task A				Task B			
System	Id	F1_TW	Δ (%)	F1_NW	Δ (%)	F1_TW	Δ (%)	F1_NW	Δ (%)
Baseline_SVC		*0.721*	–	*0.621*	–	*0.715*	–	*0.669*	–
TheNorth	2	0.809	+**12.20**	–	–	0.768	+7.41	–	–
TheNorth	1	–	–	–	–	0.772	+7.97	–	–
CHILab	1	–	–	0.774	+24.64	–	–	0.720	+7.62
MTL_AlBERTo (HS/Stereotype-Irony)		–	–	–	–	–	–	0.752	+**12.41**
MTL_ItBERT (HS/Stereotype-Irony)		0.781	+8.32	–	–	–	–	–	–
MTL_ItBERTXXL (HS/Stereotype-Irony)		–	–	–	–	0.767	+7.27	–	–
MTL_ItBERTXXL (HS/Stereotype-Irony)+Feat		–	–	0.671	+8.05	–	–	–	–
MTL_AlBERTo (HS/Stereotype-Sarcasm)		–	–	–	–	–	–	0.715	+6.87
MTL_ItBERTXXL (HS/Stereotype-Sarcasm)		0.789	+**9.43**	0.676	+8.86	–	–	–	–
MTL_ItBERTXXL (HS/Stereotype-Sarcasm)+Feat		–	–	–	–	0.749	+4.75	–	–
MTL_ItBERT (HS-Stereotype)+Feat		0.808	+**12.07**	–	–	–	–	–	–
MTL_ItBERTXXL (HS-Stereotype)		–	–	–	–	–	–	0.727	+8.67
MTL_ItBERTXXL (HS-Stereotype)+SelectedFeat		–	–	0.693	+**11.59**	0.770	+7.69	–	–

designed linguistic features seem to help the systems to go beyond the stylistic aspects, the best performance on hate speech detection in news headlines is obtained by the CHILab team exploiting the syntactic representation of text.

5. Conclusion

In this work, we investigated the linguistic and cognitive aspects involved in the explicit and implicit manifestations of abusive language in social media and news with the aim to provide a

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

solid basis for computational applications. Inspired by previous works, we examined the presence of ironic language, specific linguistic patterns, and the co-occurrence of various forms of abusive language in Italian text, exploiting the framework of the HaSpeeDe shared task.

In particular, we carried out various statistical and computational analyses that shed light, firstly, on how hate speech and stereotypes are expressed in texts from different sources, and, secondly, on how, as orthogonal dimensions of abusive language, interact between them.

The statistical analyses revealed a strong association, confirmed also at computational level, between the presence of hate speech and sarcasm in spontaneous texts such as tweets. Sarcasm is a sharp form of irony, used for mocking and ridicule a victim (Lee and Katz 1998). For its peculiarities, it proved to be adequate to express hateful messages, lowering the social cost of what has been said (Frenda *et al.* 2022).

About stereotypes, a different trend was observed between statistical and computational analyses. Indeed, the awareness of irony improves the performance of the classifier, principally in news headlines, that stand out also for its syntactic structure. This suggests that the expressions of stereotypes are characterized by specific patterns typical also of ironic language, such as the reference to secondary or indirect meanings.

Moreover, although general negative emotions and offenses appear similarly in texts containing hate speech and stereotypes, the analysis of the relevance of linguistic features shows some differences. For instance, offenses related especially to animals, physical disabilities or diversity, behaviors/morality, and general swear words are more relevant in hate speech, whereas in stereotyped messages, the offenses more significant are linked to economic and social issues, cognitive, and ethnic sphere. Stereotypes, in addition, are characterized by particular features aimed at capturing the semantic incongruity within the text, commonly used also to express irony [RQ1].

Stereotypes are conceived as oversimplified judgments shared by the members of a group that reinforce the *outgroup homogeneity* perceived as different or in contrast (Fiske 1998). Differently from hate speech, stereotypes rarely appear explicitly aggressive or offensive. For this reason, Wiegand *et al.* (2021) categorized them as types of implicit abuses that could be used to justify the hate. This explains the good performance reached in hate speech detection, when the classifier learned to recognize also stereotypes [RQ2].

However, when we process the texts only at message level, we miss contextual information. Indeed, reading the text in isolation oversimplifies how hate speech happens in reality. And, even if some texts are clearly abusive, in the other cases, context could help to give a more informed perspective to interpret them as abuses or not. In this line, as further work, we want to investigate the impact of contextual information (images, urls, conversational thread) on the resolution of implicitness in abusive language.

6. Ethical considerations

The issue faced in this paper reflects a real social problem, and we are aware of the fact that some readers could feel offended by the reported examples. Their illocutory force, in both explicit and implicit cases, is strong and reinforced by the fact that the addressed targets are entire identity groups and the offenses could touch us personally. Taking into account the sensibility of this issue, we preferred to anonymize the users' names and replaced the urls with the label URL.

We want to underline that in no way these examples reflect the opinion of the authors. The aim of this work is to create a solid foundation to theoretical debate and computational applications of prevention and detection of abusive language, encouraging academy and industry to take into account its implicitness that has severe effects like direct offenses.

Acknowledgements. The work of S. Frenda and V. Patti was partially funded by the research projects "STudying European Racial Hoaxes and sterEOTYPES" (STERHEOTYPES, under the call "Challenges for Europe" of VolksWagen Stiftung and Compagnia di San Paolo). The work of P. Rosso was partially funded by the Spanish Ministry of Science and Innovation under

the research project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media "FAKE news and HATE speech" (PGC2018-096212-B-C31) and by the Generalitat Valenciana under DeepPattern (PROMETEO/2019/121).

References

Akhtar S., **Basile V.** and **Patti V.** (2020). Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, pp. 151–154.

Basile V., **Bosco C.**, **Fersini E.**, **Nozza D.**, **Patti V.**, **Rangel-Pardo F.M.**, **Rosso P.** and **Sanguinetti**, **M.** (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*, Minneapolis, Minnesota, USA. Association for Computational Linguistics, pp. 54–63.

Basile V., **Lai M.** and **Sanguinetti M.** (2018). Long-term social media data collection at the University of Turin. In *Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy. CEUR.org.

Basile V. and **Nissim M.** (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, United States, pp. 100–107.

Bassignana E., **Basile V.** and **Patti V.** (2018). Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, vol. 2253. CEUR-WS, pp. 1–6.

Baxter J. (1997). A Bayesian/Information theoretic model of learning to learn via multiple task sampling. *Machine Learning* **28**(1), 7–39.

Bianchi C. (2021). *Hate speech: Il lato oscuro del linguaggio*. Editori Laterza.

Bosco C., **Dell'Orletta F.**, **Poletto F.**, **Sanguinetti M.** and **Tesconi M.** (2018). Overview of the EVALITA 2018 hate speech detection task. In Caselli T., Novielli N., Patti, V. and Rosso P. (eds), *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) Co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12–13, 2018*. CEUR Workshop Proceedings, vol. 2263, Turin, Italy. CEUR-WS.

Bowes A. and **Katz A.** (2011). When sarcasm stings. *Discourse Processes: A Multidisciplinary Journal* **48**(4), 215–236.

Calvin R., **Winters K.**, **Wyatt S.B.**, **Williams D.R.**, **Henderson F.C.** and **Walker E.R.** (2003). Racism and cardiovascular disease in african americans. *The American Journal of the Medical Sciences* **325**(6), 315–331.

Capozzi A.T.E., **Lai M.**, **Basile V.**, **Musto C.**, **Polignano M.**, **Poletto F.**, **Sanguinetti M.**, **Bosco C.**, **Patti V.**, **Ruffo G.**, **Semeraro G.** and **Stranisci M.** (2019). Computational linguistics against hate: Hate speech detection and visualization on social media in the "Contro L'Odio" project. In Bernardi R., Navigli R. and Semeraro G. (eds), *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13–15, 2019. CEUR Workshop Proceedings*, vol. 2481. CEUR-WS.

Carmona M.Á.Á., **Guzmán-Falcón E.**, **Montes-y-Gómez M.**, **Escalante H.J.**, **Pineda L.V.**, **Reyes-Meza V.** and **Sulayes A.R.** (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In Rosso P., Gonzalo J., Martínez R., Montalvo S. and de Albornoz J.C. (eds), *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*. CEUR Workshop Proceedings, vol. 2150. CEUR-WS, pp. 74–96.

Caselli T., **Basile V.**, **Mitrović J.**, **Kartoziya I.** and **Granitzer M.** (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In *Proceedings of The 12th Language Resources and Evaluation Conference.* European Language Resources Association, pp. 6193–6202.

Cignarella A.T., **Basile V.**, **Sanguinetti M.**, **Bosco C.**, **Benamara F.** and **Rosso P.** (2020). Multilingual irony detection with dependency syntax and neural models. In *Proceedings of the 28th International Conference on Computational Linguistics*. ACL, pp. 1346–1358.

Cignarella A.T., **Bosco C.** and **Rosso P.** (2019). Presenting TWITTIRO-UD: An Italian Twitter treebank in universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pp. 190–197.

Cignarella A.T., **Frenda S.**, **Basile V.**, **Bosco C.**, **Patti V.**, **Rosso P.** et al. (2018). Overview of the Evalita 2018 task on irony detection in Italian tweets (IronIta). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, vol. 2263. CEUR-WS, pp. 1–6.

Clarke I. and **Grieve J.** (2017). Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada. Association for Computational Linguistics, pp. 1–10.

Comandini G. and **Patti V.** (2019). An impossible dialogue! nominal utterances and populist rhetoric in an Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy. Association for Computational Linguistics, pp. 163–171.

Davidson T., **Warmsley D.**, **Macy M.** and **Weber I.** (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15–18, 2017*, vol. 11. AAAI Press, pp. 512–515.

Devlin J., **Chang M.-W.**, **Lee K.** and **Toutanova K.** (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171–4186.

Dews S. and **Winner E.** (1995). Muting the meaning a social function of irony. *Metaphor and Symbol* **10**(1), 3–19.

Dinakar K., **Reichart R.** and **Lieberman H.** (2011). Modeling the detection of textual cyberbullying. In *The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011*. AAAI Workshops, vol. WS-11-02. AAAI.

Erjavec K. and **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.** and **Rosso P.** (2020). AMI @ EVALITA2020: Automatic misogyny identification. In Basile V., Croce D., Maro M.D. and Passaro L.C. (eds), *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*. CEUR Workshop Proceedings, vol. 2765. CEUR-WS.

Fersini E., **Rosso P.** and **Anzovino M.** (2018). Overview of the task on automatic misogyny identification at IberEval 2018. In Rosso P., Gonzalo J., Martínez R., Montalvo S. and de Albornoz J.C. (eds), *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*. CEUR Workshop Proceedings, vol. 2150. CEUR-WS, pp. 214–228.

Fiske S.T. (1998). Stereotyping, prejudice, and discrimination. *The Handbook of Social Psychology* **2**(4), 357–411.

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

Frenda S. (2018). The role of sarcasm in hate speech: A multilingual perspective. In *Proceedings of Doctoral Symposium at SEPLN 2018*. CEUR-WS.

Frenda S., **Banerjee S.**, **Rosso P.** and **Patti V.** (2020). Do linguistic features help deep learning? The case of aggressiveness in Mexican Tweets. *Computación y Sistemas* **24**(2), 633–643.

Frenda S., **Cignarella A.T.**, **Basile V.**, **Bosco C.**, **Patti V.** and **Rosso P.** (2022). The unbearable hurtfulness of sarcasm. *Expert Systems with Applications* **193**, 116398.

Fulper R., **Ciampaglia G.L.**, **Ferrara E.**, **Ahn Y.**, **Flammini A.**, **Menczer F.**, **Lewis B.** and **Rowe K.** (2014). Misogynistic language on twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on ChASM*.

Gambino G. and **Pirrone R.** (2020). CHILab @ HaSpeeDe 2: Enhancing hate speech detection with part-of-speech tagging. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Garavelli B.M. (1997). *Manuale di retorica*. Bompiani Milan.

Giora R., **Givoni S.** and **Fein O.** (2015). Defaultness reigns: The case of sarcasm. *Metaphor and Symbol* **30**(4), 290–313.

Grice H.P. (1975). Logic and conversation. In *Speech Acts*. Brill, pp. 41–58.

Hernández Farías D.I., **Patti V.** and **Rosso P.** (2016). Irony detection in Twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)* **16**(3), 19.

Joshi A., **Sharma V.** and **Bhattacharyya P.** (2015). Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, pp. 757–762.

Jurgens D., **Hemphill L.** and **Chandrasekharan E.** (2019). A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 3658–3666.

Karoui J., **Farah B.**, **Moriceau V.**, **Patti V.**, **Bosco C.** and **Aussenac-Gilles N.** (2017). Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, pp. 262–272.

Kumar R., **Ojha A.K.**, **Malmasi S.** and **Zampieri M.** (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 1–11.

Kumar R., **Ojha A.K.**, **Malmasi S.** and **Zampieri M.** (2020). Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France. European Language Resources Association (ELRA), pp. 1–5.

Lapidot-Lefler N. and **Barak A.** (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* **28**(2), 434–443.

Lavergne E., **Saini R.**, **Kovács G.** and **Murphy K.** (2020). Thenorth@ HaSpeeDe 2: BERT-based language model fine-tuning for italian hate speech detection. In *7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020*, vol. 2765. CEUR-WS.

Lee C.J. and **Katz A.N.** (1998). The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol* **13**(1), 1–15.

Lemmens J., **Markov I.** and **Daelemans W.** (2021). Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pp. 7–16.

MacAvaney S., **Yao H.-R.**, **Yang E.**, **Russell K.**, **Goharian N.** and **Frieder O.** (2019). Hate speech detection: Challenges and solutions. *PloS One* **14**(8), e0221152.

Mandl T., **Modha S.**, **Kumar M A.** and **Chakravarthi B.R.** (2020). Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In Forum for Information Retrieval Evaluation, pp. 29–32.

Mohammad S.M. and **Turney P.D.** (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* **29**(3), 436–465.

Müller K. and **Schwarz C.** (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* **19**(4), 2131–2167.

Nadal K.L., **Griffin K.E.**, **Wong Y.**, **Hamit S.** and **Rasmus M.** (2014). The impact of racial microaggressions on mental health: Counseling implications for clients of color. *Journal of Counseling & Development* **92**(1), 57–66.

Nobata C., **Tetreault J.**, **Thomas A.**, **Mehdad Y.** and **Chang Y.** (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 145–153.

Ortiz Suárez P.J., **Sagot B.** and **Romary L.** (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Mannheim*. Leibniz-Institut für Deutsche Sprache, pp. 9–16.

Pamungkas E.W., **Basile V.** and **Patti V.** (2020). Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. In *Information Processing & Management*, **57**(6), 102360.

Pan H., **Lin Z.**, **Fu P.** and **Wang W.** (2020). Modeling the incongruity between sentence snippets for sarcasm detection. In *24th European Conference on Artificial Intelligence, 29 August–8 September 2020, Santiago de Compostela, Spain – Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*.

Pexman P.M. and **Olineck K.M.** (2002). Does sarcasm always sting? Investigating the impact of ironic insults and ironic compliments. *Discourse Processes* **33**(3), 199–217.

Plutchik R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist* **89**(4), 344–350.

Plutchik R. and **Kellerman H.** (2013). *Theories of Emotion*, vol. **1**. New York: Academic Press.

Poletto F., **Basile V.**, **Sanguinetti M.**, **Bosco C.** and **Patti V.** (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation* **55**, 477–523.

Polignano M., **Basile P.**, **de Gemmis M.**, **Semeraro G.** and **Basile V.** (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on Tweets. In *CLiC-it*.

Riloff E., **Qadir A.**, **Surve P.**, **De Silva L.**, **Gilbert N.** and **Huang R.** (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 704–714.

Rösner L. and **Krämer N.C.** (2016). Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media + Society* **2**(3), 1–13.

Röttger P., **Vidgen B.**, **Nguyen D.**, **Waseem Z.**, **Margetts H.** and **Pierrehumbert J.** (2021). HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics, pp. 41–58.

Sanguinetti M., **Comandini G.**, **Di Nuovo E.**, **Frenda S.**, **Stranisci M.A.**, **Bosco C.**, **Caselli T.**, **Patti V.** and **Russo I.** (2020). HaSpeeDe 2@ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, vol. 2765. CEUR-WS, pp. 1–9.

Sanguinetti M., **Poletto F.**, **Bosco C.**, **Patti V.** and **Stranisci M.** (2018). An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA), pp. 2798–2895.

Schmidt A. and **Wiegand M.** (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10.

Talat Z., **Thorne J.** and **Bingel J.** (2018). *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*. Cham: Springer International Publishing, pp. 29–55.

Taulé M., **Ariza A.**, **Nofre M.**, **Amigó E.** and **Rosso P.** (2021). Overview of DETOXIS at IberLEF 2021: DEtection of TOXicity in comments In Spanish. *Procesamiento del Lenguaje Natural* **67**, 209–221.

Tiedemann J. and **Nygaard L.** (2003). OPUS-an open source parallel corpus. In *Proceedings of the 13th Nordic Conference on Computational Linguistics (NODALIDA). University of Iceland, Reykjavik*.

Van Aken B., **Risch J.**, **Krestel R.** and **Löser A.** (2018). Challenges for toxic comment classification: An in-depth error analysis. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW 2)*. Association for Computational Linguistics, Belgium, pp. 33–42.

https://doi.org/10.1017/S1351324922000316 Published online by Cambridge University Press

Vidgen B. and **Derczynski L.** (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE* **15**(12), 1–32.

Vidgen B., **Thrush T.**, **Waseem Z.** and **Kiela D.** (2021). Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics, pp. 1667–1682.

Waseem Z., **Davidson T.**, **Warmsley D.** and **Weber I.** (2017). Understanding abuse: A typology of abusive language detection subtasks. arXiv preprint arXiv:1705.09899.

Waseem Z. and **Hovy D.** (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, San Diego, California. Association for Computational Linguistics, pp. 88–93.

Wiegand M., **Ruppenhofer J.** and **Eder E.** (2021). Implicitly abusive language–What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 576–587.

Wiegand M., **Ruppenhofer J.** and **Kleinbauer T.** (2019). Detection of abusive language: The problem of biased datasets. In *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. 602–608.

Wiegand M., **Ruppenhofer J.**, **Schmidt A.** and **Greenberg C.** (2018). Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 11046–1056.

Williams M.L., **Burnap P.**, **Javed A.**, **Liu H.** and **Ozalp S.** (2020). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology* **60**(1), 93–117.

Zampieri M., **Malmasi S.**, **Nakov P.**, **Rosenthal S.**, **Farra N.** and **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*, Minneapolis, Minnesota, USA. Association for Computational Linguistics, pp. 75–86.

Cite this article: Frenda S, Patti V and Rosso P. Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering* https://doi.org/10.1017/S1351324922000316