Quality & Quantity
https://doi.org/10.1007/s11135-022-01494-7

An Italian lexical resource for incivility detection in online discourses

Alice Tontodimamma[1] · Lara Fontanella[2] · Stefano Anzani[1,3] · Valerio Basile[4]

Accepted: 5 July 2022

Abstract
The exponential growth of social media has brought an increasing propagation of online hostile communication and vitriolic discourses, and social media have become a fertile ground for heated discussions that frequently result in the use of insulting and offensive language. Lexical resources containing specific negative words have been widely employed to detect uncivil communication. This paper describes the development and implementation of an innovative resource, namely the Revised HurtLex Lexicon, in which every headword is annotated with an offensiveness level score. The starting point is HurtLex, a multilingual lexicon of hate words. Concentrating on the Italian entries, we revised the terms in HurtLex and derived an offensive score for each lexical item by applying an Item Response Theory model to the ratings provided by a large number of annotators. This resource can be used as part of a lexicon-based approach to track offensive and hateful content. Our work comprises an evaluation of the Revised HurtLex lexicon.

Keywords Uncivil speech detection · Lexicon-based approach · Sentiment analysis · Item response theory model · Raters' consistency

Alice Tontodimamma, Lara Fontanella, Stefano Anzani and Valerio Basile have contributed equally to this work.

✉ Lara Fontanella
lara.fontanella@unich.it

Alice Tontodimamma
alice.tontodimamma@unich.it

Stefano Anzani
stefano.anzani@unich.it

Valerio Basile
valerio.basile@unito.it

[1] Department of Neuroscience, Imaging and Clinical Sciences, G. d'Annunzio University of Chieti-Pescara, Via dei Vestini 33, 6610 Chieti, Italy

[2] Department of Legal and Social Sciences, G. d'Annunzio University of Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy

[3] Center for Advanced Studies and Technology (CAST), G. d'Annunzio University of Chieti-Pescara, Via dei Vestini 31, 6610 Chieti, Italy

[4] Computer Science Department, University of Turin, Via Verdi 8, 10124 Turin, Italy

Mathematics Subject Classification 91F20 · 62H25

1 Introduction

Over the past years, the exponential growth of social media has increased the propagation of hostile online communication and vitriolic discourses. While the eased access to digital public spheres offers new opportunities for active participation in public debates, it also provides new possibilities for norm transgressing forms of online engagement, and incivility spread more rapidly and widely than ever before. In fact, in face-to-face interactions there are strong social norms likely to be observed. On the other hand, in occasional Internet-mediated interactions, unknown strangers are basically 'invisible', and their feelings and sensitiveness can hardly be perceived. Therefore, online interactions are more likely to include aggressive comments, harsh critiques, hate speech and harassment (Sabatini and Sarracino 2019).

Scholars have defined incivility as the communication of disagreement combined with a dismissive, disrespectful, aggressive, or hostile tone (Coe et al. 2014). Incivility can be understood as norm-transgressing communication. In our work, we adhere to the distinction by Frischlich et al. (2021): when the transgression is related to norms regulating interpersonal communication, we refer to 'offensive speech', while when the violation relates to intergroup communication norms, we consider that incivility falls into the realm of 'hate speech'. Under this framework, incivility as 'offensive speech' is strictly linked to the flaming phenomenon, which broadly consists of any aggressive or hostile interpersonal communication occurring via computer-mediated channels (O'Sullivan and Flanagin 2003). On the other hand, hate speech can be more strictly defined as any bias-motivated, hostile, malicious communication aimed at a person or a group of people because of some of their actual or perceived innate characteristics such as race, colour, ethnicity, gender, sexual orientation, nationality, religion (Cohen-Almagor 2011).

Various rhetorical and stylistic elements have been labelled as uncivil, including name-calling, slurs, vulgarity, profanity, and derogatory speech. Attempts to study and define uncivil communication or to make a linguistic description to identify its characteristics begin from the lexical level. In fact, the most obvious and most accessible departure point for defining and studying offensive and hate speech is the detection of abusive words that cause pain being derogatory in nature. These are the worst words you could use, especially if you are part of a group with power over another group that, because of minority status or history of discrimination, has less power (Faloppa 2020). The presence of offensive and threatening language in a text could represent a feature to identify such a text as uncivil. To this end, many authors (see Schmidt and Wiegand 2017, for a comprehensive review) rely on lexical resources containing specific negative words (e.g. slurs, insults, swearing expressions). The most of lexical resources to detect toxic, hateful, offensive or abusive contents are proposed for English and, until recently, little attention has been paid to the creation of Italian lexicons.

The main contribution of this work is the development of an innovative computational instrument for incivility detection: a lexicon in which every headword is annotated with an offensiveness level score that could be helpful to track down offensive and hateful expressions. The starting point of our work is HurtLex, a multilingual lexical computational resource for hate speech detection developed by Bassignana et al. (2018). Focusing

on Italian words, we revised the terms in HurtLex and derived an offensive score for each lexicon entry by using an Item Response Theory (IRT) model (de Ayala 2009).

In order to understand the association of offensiveness and sentiment polarity, we decided to exploit the Revised HurtLex lexicon in conjunction with a lexicon-based Sentiment Analysis approach (Taboada et al. 2011). In addition, to evaluate the performance of the revised lexicon on a downstream task, we performed a supervised classification experiment.

The rest of this paper is organised as follows: in Sect. 2 we outline lexical resources used for offensive and hate speech detection. The materials and methods used to create the lexicon are detailed in Sect. 3. In particular, Sect. 3.3 provides the IRT model results, in terms of raters' consistency assessment and words' offensiveness level estimation. Section 4 gives an evaluation of the Revised HurtLex lexicon for uncivil discourse detection. Finally, in Sect. 5, our conclusions are given, and some future projects are outlined.

2 Lexical resources for offensive and hate speech detection

In the framework of sentiment analysis and opinion mining, there are three main ways of building sentiment lexicons: hand-craft elaboration, dictionary-based approaches and corpus-based approaches (Almatarneh and Gamallo 2018). In dictionary-based methods, an automatic expansion from an initial list of seed words is performed. The new polarity entries are mainly based on the synonyms and antonyms of external resources, such as a thesaurus. Corpus-based approaches also make use of a list of seed sentiment words to find other sentiment words and their polarity from the given corpus, leading to the construction of domain-dependent polarity lexicons.

Several approaches have been proposed to create computer-based lexical resources designed to identify hate speech or abusive content. Primarily, some approaches make direct use of lists of derogatory and abusive words posted on the web. Schmidt and Wiegand (2017) provide several references to publicly available lists that consist of general hate-related terms or that are specialized towards a particular subtype of hate speech, such as ethnic slurs, LGBT slang terms, or words with a negative connotation towards disabled people.

Other approaches incorporate lists explicitly created for the task at hand (see Schmidt and Wiegand 2017; Poletto et al. 2021, for a review).

Among these, the 'Insulting and Abusing Language Dictionary', manually compiled by Razavi et al. (2010), which includes words and phrases used with different degrees of manifestation of hatred. Each entry was initially assigned a weight in the range of 1–5 on the basis of its potential impact on the classification of the containing context. Then, adaptive learning on training data was performed to obtain the final weights.

Gitari et al. (2015) compiled a list of hate verbs that condone or encourage acts of violence. Wiegand et al. (2018) address the problem of detecting abusive words among a set of negative polar expressions. Their starting list contains negative words drawn from the Subjectivity Lexicon (Wilson et al. 2005) along with some prototypical abusive words frequently occurring in the word lists mentioned in Schmidt and Wiegand (2017). Those initial negative words were binary classified as abusive or non-abusive by five annotators via crowdsourcing. An agreement of at least four out of the five raters judged was required to include the word in the abusive base lexicon. This base lexicon was then expanded through automatic classification employing information from both corpora and lexical resources.

Vargas et al. (2021) propose a contextual and cross-lingual offensive lexicon (MOL—Multilingual Offensive Lexicon) composed of 1,000 explicit and implicit offensive and swearing expressions. Offensive terms are defined as term or expression that intends to undermine or disparage any of the following social aspects: moral, appearance, physical, psychological health, sexual behaviour and orientation, intellectual, economic, religious and political aspects. On the other hand, swearing consists of terms or expressions used to convey hateful opinions, with highly aggressive value and great potential to generate negative reactions to the interlocutor. The terms were extracted manually from a corpus of hate speech and offensive comments on social media. Three different annotators carried out the annotation process.

As for offensive Italian lexicons, a list of taboo words and expressions has been proposed by Maisto et al. (2017) for mining offensive language on social media. Italian offensive and hateful words can also be found in HurtLex (Bassignana et al. 2018), a multilingual lexicon built semi-automatically from the originally handcrafted Italian lexicon 'Words as Weapons' (De Mauro 2016). The words in HurtLex are divided into 17 overlapping categories and marked for the presence of stereotypes. HurtLex also excels with additional linguistic information (parts of speech, lexicographic definitions) for its lemmas.

3 The revised HurtLex lexicon: materials and methods

As previously stated, the central contribution of this work concerns the construction of a lexical resource, in which every headword is annotated with an offensiveness level score. Materials and methods used to create this lexical resource are detailed in this section, along with the results obtained by implementing the IRT model to assess the intra-rater consistency and estimate the offensiveness level for the lexicon entries.

3.1 Materials

To build a computational lexical resource for offensive and hateful content detection, we used as a starting point the Italian version of Hurtlex. In particular, the building of this computational resource involved the following steps.

The first step consists of removing non-offensive terms and adding potentially offensive terms. Specifically, from the original list, we omitted terms that cannot be used as 'weapons' (e.g., '*adolescente*','*accademico*', '*benpensante*', '*battuta*', '*ideologia*', '*scarabbocchiare*'), terms not used in Italian (e.g., '*balladeer*', '*blackball*', '*darkey*', '*finagler*', '*hyaena*', '*sniveller*', '*sus*', '*wuss*') and meaningless terms (e.g., '*agdgadu*', '*allevatore impianto*', '*capo-decina*', '*gallus gallus domesticus*', '*pandar*', '*ybris*'). We added new words, looking out for synonyms of terms already present in the original version, consulting various web pages where insults are listed and examining the list of 'bad words'.[1] Other potentially offensive words were manually selected from the vocabulary of multiple corpora. In particular, we considered: the Italian Hate Speech Corpus (IHSC, Sanguinetti et al. 2018); the AMI Corpus on Misogyny[2] (Fersini et al. 2020); a corpus of our own, created by downloading, from Facebook, YouTube, and Twitter, 433,003 comments on migrants,

[1] lista_badwords.txt, https://github.com/napolux/paroleitaliane/tree/master/paroleitaliane.

[2] https://amievalita2018.wordpress.com/.

women, Roma people, and the LGBTQI+ community. These corpora were cleaned, tokenized and their vocabulary was extracted. Trainees read through the vocabulary looking for potentially offensive terms. They marked all the words that they considered offensive, and we added to our dictionary all words that were not already in there. The inclusion/removal process was carried out by six trainees who were engaged in an internship program and was revised by the authors. In the end, 2140 terms from the original HurtLex were not included in the revised new version and 3419 new terms were added.

In the second step for each term, we included all the grammatical forms in which they occur. We added singular (masculine and feminine) and plural (masculine and feminine) forms for each noun and adjective. As verbs, we decided to include (in addition to the base form already present in HurtLex) the forms and tenses most used to offend. In the third step, we reviewed the categories. For each new entry, its categorisation was included, using the existing HurtLex categories and adding to those suggested by De Mauro (2016), and included in the original version of HurtLex, four new categories: nouns and adjectives used to offend, insult, or denigrate women; words used to intimidate other people; words denoting insults connected to political matters; nouns related to illnesses and diseases.

Finally, in the last step, we graded the offensiveness of each entry. To this end, since determining a given expression's offensiveness is a highly subjective matter, we decided to create a measurement instrument to be administered to a group of selected respondents.

Researchers in many disciplines, including computational linguistics, often rely on rating scales when manually annotating data. In particular, in building a sentiment or emotional lexicon, the annotations for any given item can be gathered by collecting individual ratings generally on a 5-point or 9-point Likert scale (see BeersFägersten 2007; Warriner et al. 2013; Buechel and Hahn 2018; Zhao et al. 2019, among others). The individual assessments from multiple raters are usually averaged to obtain a real-valued score for each item.

In our work, we adopted a rating scale annotation such that each word was to be assessed on a 5-point Likert scale, plus a zero rating for any term that the respondent considered unoffensive. Our lexicon contains nouns, adjectives, verbs, adverbs, and interjections. In order to reduce the number of words that the raters would have to evaluate, from the starting 2639 base forms for nouns and adjectives we kept only their singular form, using either the masculine or feminine. We also excluded all the words that could be considered synonyms of other words in the list, since we could simply assign them the same score of the corresponding synonym.

3.2 Methods

Different authors (see Kiritchenko and Mohammad 2017; Poletto et al. 2019, and references therein) point out how rating scales present several challenges, such as the difficulty in maintaining inter- and intra-annotator consistency.

To overcome this consistency issue, instead of simply taking the average of the individual ratings, we rely on an item factor analytic model suitable for ordinal responses. In particular, to derive an offensiveness index for each headword included in the measurement instrument we explicitly model the item response probability trough a unidimensional Graded Response Model (GRM) (Samejima 1969) where the latent trait is related to the words while the so-called item parameters are related to raters. More specifically, we assume that the score assigned to a chosen term depends on its intrinsic latent level of offensiveness and on some respondent's parameters. The respondent's

parameters represent the weight of the rater's evaluation in the composite index of offensiveness and the threshold values between consecutive categories along his/her continuous evaluation scale.

Formally, denoting by X_{ij} the score assigned by rater $j = 1, \ldots, N$ to word $i = 1, \ldots, K$, according to the two-parameter normal ogive (2-PNO) formulation of the GRM model, the probability that the score will fall into category $c = 1, \ldots, C$ of the ordered response scale is given by

$$P(X_{ij} = c|\theta_i, \lambda_j, \boldsymbol{\gamma}_j) = \Phi\big(\lambda_j\theta_i - \gamma_{j,c-1}\big) - \Phi\big(\lambda_j\theta_i - \gamma_{j,c}\big). \tag{1}$$

Here, Φ is the standard normal distribution, θ_i denotes the level of offensiveness of word i, λ_j is the factor loading, or discrimination parameter, for rater j, and finally $\boldsymbol{\gamma}_j = (\gamma_{j,1} \cdots, \gamma_{j,C-1})'$ is the vector of ordered threshold parameters $-\infty \leq \gamma_{j,1} \leq \cdots \leq \gamma_{j,C-1} \leq \infty$.

For computational purposes, this model is developed in a Bayesian framework by exploiting the alternative formulation (Skrondal and Rabe-Hesketh 2005) which defines a set of underlying continuous variables $\mathbf{Z}_i = (Z_{i,1}, Z_{i,2}, \ldots, Z_{i,N})'$ that can be expressed in terms of a linear transformation of the latent trait θ through the discrimination parameters $\lambda = (\lambda_1, \ldots, \lambda_N)'$:

$$\mathbf{Z}_i = \lambda\theta_i + \boldsymbol{\epsilon}_i \tag{2}$$

where $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{I})$, with $\mathbf{I}$ denoting the identity matrix. Each categorical rating X_{ij} is linked to the underlying continuous response Z_{ij} through the following threshold model:

$$X_{ij} = c \quad if \quad \gamma_{j,c-1} \leq Z_{ij} \leq \gamma_{j,c}, \quad for \quad c = 1, \ldots, C \tag{3}$$

where $\gamma_{j,0} = -\infty$ and $\gamma_{j,C} = \infty$.

For parameter estimation, we adopt Markov Chain Monte Carlo (MCMC) simulation techniques. In our model specification, to address location and scale indeterminacy, we assume a zero mean and unit variance latent trait. More specifically, in the prior specifications, we assume that all word parameters are independent and identically distributed samples from a normal distribution, that is $\forall\, i : \theta_i \sim \mathcal{N}(0, 1)$. We define a normal prior for the discrimination parameters, $\forall j : \lambda_j \sim \mathcal{N}(0, 1)$ and a uniform prior distribution for the threshold parameters $\gamma_{j,c}$ truncated to the region $\{\gamma_{j,c} \in \mathcal{R}, \gamma_{j,c-1} \leq \gamma_{j,c} \leq \gamma_{j,c+1}\}$, $c = 1, \ldots, C-1$, $\forall j$, to take account of the order constraints (Albert and Chib 1993). The full conditional of most parameters can be specified in closed form which allows for a Gibbs sampler although Metropolis–Hastings steps are required to sample the threshold parameters. In particular to simulate the thresholds we exploit the Cowles' algorithm (Cowles 1996).

Our formulation of the IRT model differs from conventional IRT approaches where individuals are treated as replications and the evaluated instances are treated as items. Here, on the other hand, the evaluators are treated like items so that the item parameters are taken to represent differences between raters' decision criteria. Baldwin et al. (2009) show how this alternative perspective can help in identifying differences between raters. The two parameter formulation of the GRM naturally takes into account the issue of inter-annotator consistency by allowing different discrimination parameters for the raters. In addition, we exploit the GRM to evaluate the intra-annotator consistency in two temporally distinct annotation procedures. In particular, we assess the differences in the raters' discrimination parameters and threshold vectors.

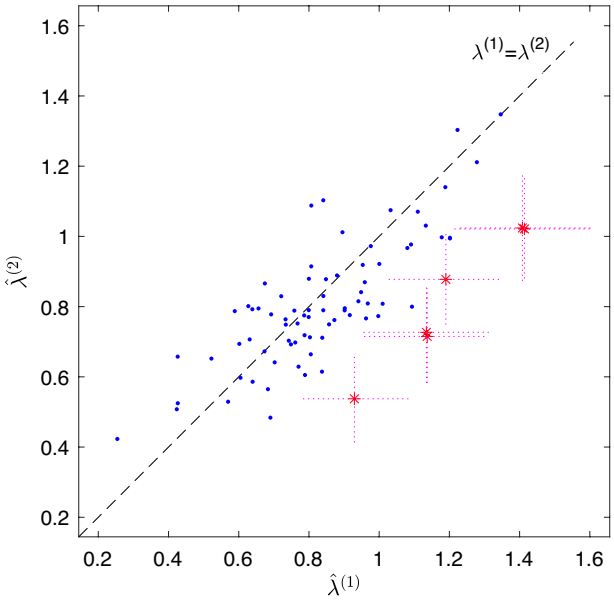

Fig. 1 Scatter plot of the discrimination parameter posterior estimates for the two temporally separated annotation procedures. The red asterisks represent the raters for which the hypothesis $\lambda_j^{(1)} = \lambda_j^{(2)}$ is not verified. The dashed lines depict the 99% credible interval

3.3 IRT model results

For the manual annotation, out of the 2639 base forms for nouns and adjectives, we selected 1238 terms. The selection was made manually, and we left out all the words that could be considered synonyms of other words already included in the list, with the idea of assigning them the same score obtained through the rating process. The 1238 words were partitioned in two separate lists administered to the evaluators two weeks apart. The scoring process was carried out by 81 raters (19 males, 62 females, aged between twenty and fifty-three). The words with at least 25% of zero ratings (66 words), were removed from the lexicon. For the remaining words, the zero ratings were assigned level one, the lowest level of offensiveness. For ninety of the words, there was at least one missing datum, where a rating had not been assigned by one or more of the evaluators. Assuming a missing at random mechanism, multiple imputation was carried out exploiting the posterior predictive distribution.

3.3.1 Intra-raters' consistency

To assess intra-annotator consistency, we compare the posterior estimates of the discrimination parameters and the thresholds for the ratings obtained in the two temporally separated annotation procedures. More in details, for each rater we verify the hypothesis of equal parameters by considering whether the corresponding 99% credible intervals overlap. As for the discrimination parameters, the hypothesis $\lambda_j^{(1)} = \lambda_j^{(2)}$ is not verified for 6 out of the 81 annotators. Figure 1 represents the scatter plot of the posterior estimates for the two annotation procedures along with their 99% credible intervals. The coefficients for which the hypothesis of equality in the two annotation procedures does not hold are depicted in red.

To assess the differences in the threshold structure, for each rater we consider the location value ι_j, determined by the mean of the rater thresholds $\boldsymbol{\gamma}_j = (\gamma_{j,1}, \ldots, \gamma_{j,C-1})'$). This

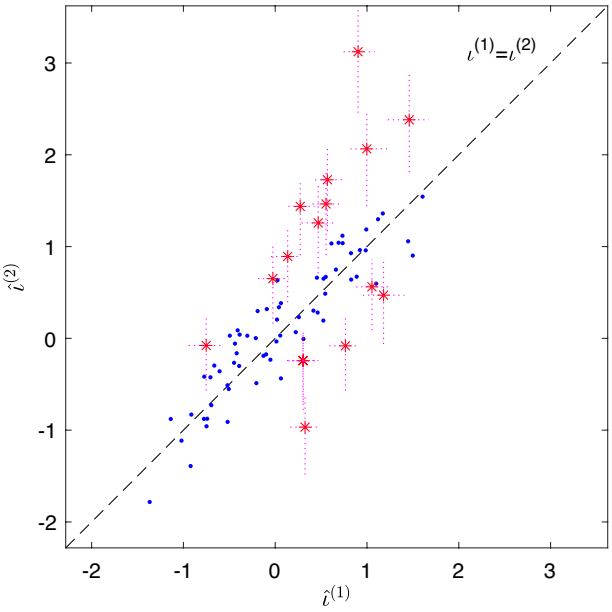

Fig. 2 Scatter plot of the location parameter posterior estimates for the two temporally separated annotation procedures. The red asterisks represent the raters for which the hypothesis $\iota_j^{(1)} = \iota_j^{(2)}$ is not verified. The dashed lines depict the 99% credible interval

mean value can be interpreted as rater's susceptibility to offensive language: higher location values correspond to lower levels of susceptibility. The hypothesis $\iota_j^{(1)} = \iota_j^{(2)}$ is rejected for 16 annotators (see Fig. 2). As can be noted from Fig. 3, the correlation between the estimated thresholds at the two temporal annotation procedures, $\hat{\gamma}_{j,c}^{(1)}$ and $\hat{\gamma}_{j,c}^{(2)}$, $c = 1, \ldots, 4$, increases when removing the thresholds of those 16 raters. The final dataset consists of the ratings provided by the 61 annotators for which the intra-raters' consistency holds. It is worth noting that for two raters the null hypothesis is rejected for both the discrimination parameters and the mean threshold values.

3.3.2 Offensiveness level estimate

As stated above, the two-parameter formulation of the GRM model enables us to estimate raters' discrimination parameters and thresholds along with the rating for each entry included in the measurement instrument.

Figure 4 shows the posterior estimates of the discrimination parameters along with their 99% credible interval. The discrimination parameter can be interpreted as the regression coefficients of the continuous variables, underlying the categorical ratings, with respect to the latent trait representing word offensiveness. All the coefficients are significantly different from zero and positive, corroborating inter-raters' concordance.

Figure 5 displays the estimated thresholds for all the raters. The red asterisks represent the mean value of the estimated thresholds for each rater and can be interpreted as the rater's susceptibility level concerning offensiveness. As can be seen, the distances between consecutive thresholds are not uniform, neither between the raters nor for each rater, and this justifies the use of an item factor analytic model that allows to take into account the different subjective rating systems. Finally, Fig. 6 shows the posterior estimates for the latent trait scores, representing the offensiveness level of each words included in the measurement instruments. As stated above, for identifiability purposes,

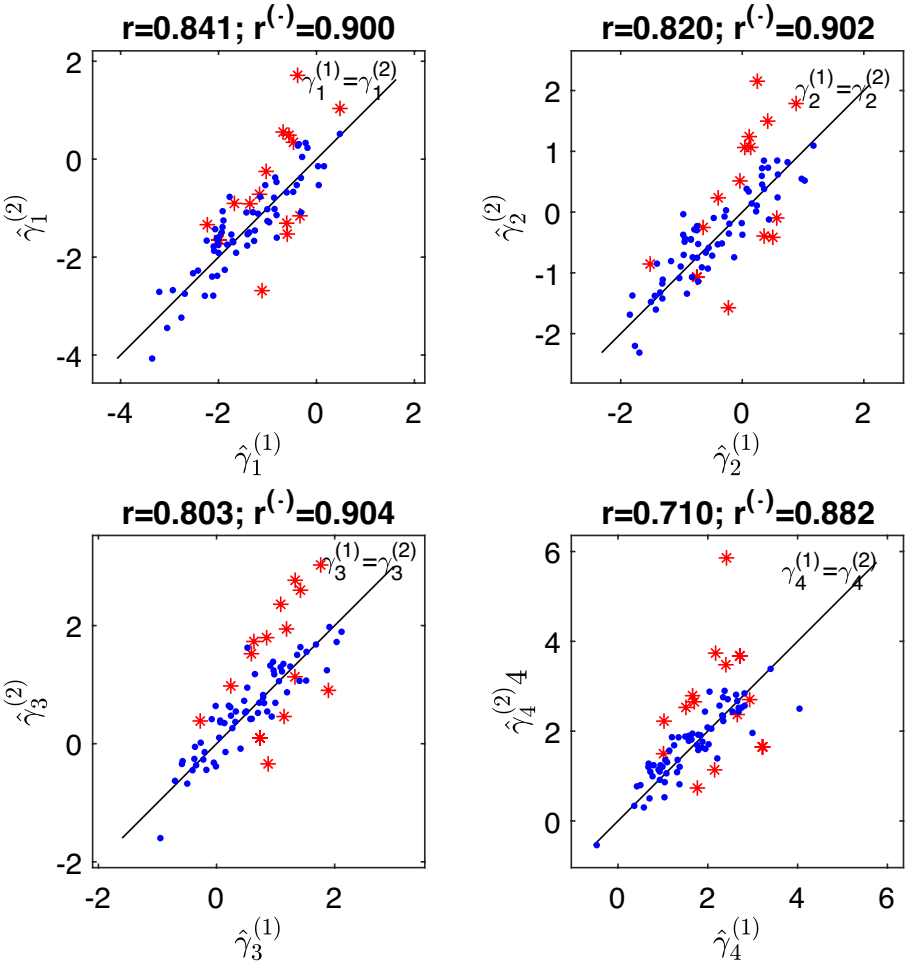

Fig. 3 Scatterplot of the thresholds posterior estimates for the two temporally separated annotation procedures. The red asterisks represent the raters for which the hypothesis $\iota_j^{(1)} = \iota_j^{(2)}$ is not verified. *r*: correlation coefficients for all the raters; $r^{(\cdot)}$ correlation coefficients for the raters for which the hypothesis $\iota_j^{(1)} = \iota_j^{(2)}$ *holds*

the mean and the standard deviation of this latent trait have been fixed to 0 and 1, respectively. As it can be noted, some words have been assigned a very high offensiveness level. Those words are mainly sexist derogatory slurs.

Given the estimated offensiveness levels for the term in the measurement instruments, we assigned an offensiveness degree to all the words in the revised HurtLex excluded from the rating process. To this purpose, we first exploited the association between words, given that the excluded terms were synonyms of terms included and, then, we assigned the same score to words with the same root. To obtain a final score, we rescaled the offensiveness level in the range from 1 to 5. The result is a lexicon in which each entry is provided with an offensiveness rating. It is worth noting that we decided to score as 6 offensive interjections, and as 5 offensive verbs, not included in the measurement instruments.

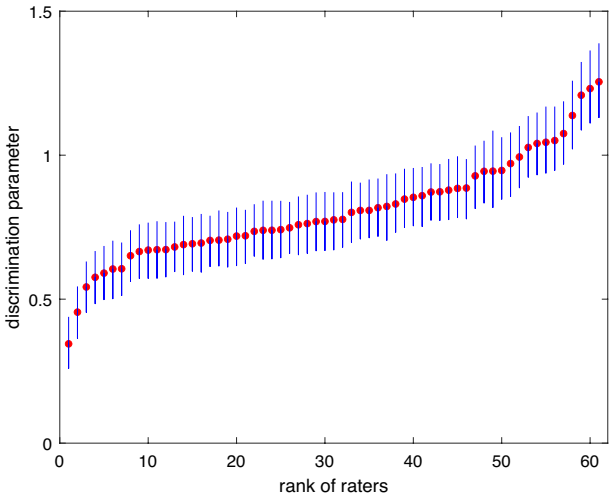

Fig. 4 Caterpillar plot of the discrimination parameter posterior estimates. The red dots represent the posterior estimates of λ_j, $j = 1, \ldots, N$. The lines depict the 99% credible interval

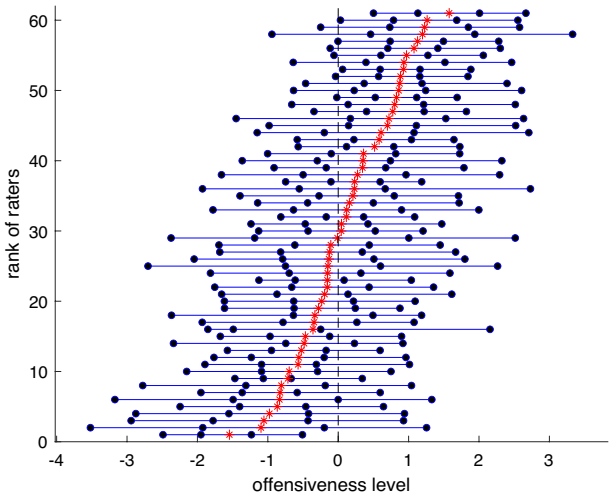

Fig. 5 Raters' map of the threshold posterior estimates (*blue dots*). The red asterisks represent the posterior estimates of ι_j, $j = 1, \ldots, N$

In total the Revised HurtLex lexicon contains 7920 words including all inflected forms: 7024 nouns and adjectives (2639 in the base form); 576 verbs and 320 interjections. As verbs, we decided to include (in addition to the base form already present in HurtLex) the forms and tenses most used to offend. The Revised HurtLex lexicon is made publicly available for download on github.[3]

[3] https://github.com/valeriobasile/hurtlex.

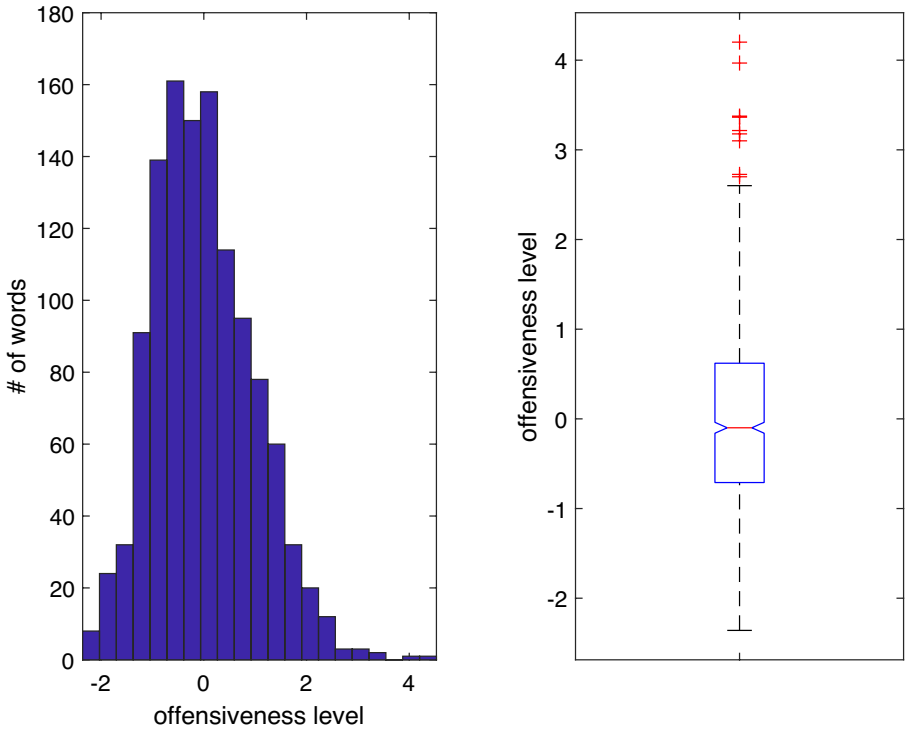

Fig. 6 Representation of the latent trait score posterior estimates

4 Detection of uncivic comments online

Usually, the semantics of hate speech contains a strong negative sentiment tendency. In order to understand the association of offensiveness and sentiment polarity, we decided to exploit the Revised HurtLex lexicon in conjunction with a lexicon-based Sentiment Analysis approach (Taboada et al. 2011) and compared comments' offensiveness level with the scores obtained using different affective lexica.

On the other hand, to explore the usefulness of the proposed lexical resource for hate speech detection, we performed a supervised classification experiment. Section 4.1 describes the textual datasets exploited in the analysis, in Sect. 4.2 we present the Sentiment Analysis results, and, finally, in Sect. 4.3 we provide the results of the supervised classification task.

4.1 The analysed corpora

The first corpus contains comments relating to the decision taken by Italian President Sergio Mattarella, on 27 May 2018, not to accept the appointment of Paolo Savona as Economy and Finance minister in the government proposed by the Northern League and the 5 Star Movement, a decision which polarised public opinion. The dispute was particularly heated on social networks, where the numerous supporters of the Head of State were

matched by equal numbers who posted insults and messages filled with hate. The corpus was assembled from a selection of comments extracted from Facebook, Youtube and Twitter. Different software was used to download texts. The comments to the post relating to the speech of May 27 by the President of the Republic on the unofficial page of Mattarella were downloaded through the NodeXL[4] software. The tweets' scraping was done through the Socialgrabber[5] software, using the following keywords: *sergiomattarella*, *matarella*, *presidentedellarepubblica*, *impeachment*. Finally, a specific Python script was written to download the comments to some videos of Mattarella's speech on YouTube. The dataset was annotated by three graduates as part of their master's thesis work on hate speech, using binary annotation: comments were marked with 1 if they contained an example of hate speech and with 0 if they did not. The annotated dataset[6] comprises 3094 textual documents, out of which 1115 were classified as hate speech.

The second corpus comprises 3967 comments downloaded from Facebook and released during the sixth evaluation campaign EVALITA 2018, specifically for the task of Hate Speech Detection (Bosco et al. 2018). The analysed corpus comprises 2058 comments classified as containing hate speech.

The third corpus comprises 3990 comments downloaded from Twitter, containing hate speech directed against migrants (Sanguinetti et al. 2018). The tweets are classified as containing hate speech and also as containing offensiveness, which, if present, can be distinguished, based on the extent of the offense, as either weak or strong. This corpus comprises 1284 comments classified as containing hate speech.

4.2 Sentiment polarity and Offensiveness level

To evaluate the association between comments' sentiment polarity and offensiveness level, we adopt a lexicon-based approach. Lexicon-based Sentiment analysis sets out to determine the orientation of a text by using dictionaries known as affective lexicons or affection lexicons, which specify the semantic orientation of words (also known as word polarity). The use of such affective lexicons enables the classifier to assign, a priori, a positive or negative polarity to the terms used in a given corpus. The whole text's polarity grading derives from the sum of the polarity gradings assigned to the entities present in the text under examination. Along the same lines, we exploit the scores assigned to the terms in the Revised HurtLex lexicon to derive a score of offensiveness for each textual document. We compared the results of the approach based on our lexical resource with the scores obtained through Sentiment analysis. In particular, for sentiment analysis, we made use of three different affective Italian lexicons, namely Sentix (Basile and Nissim 2013), the Morphologically-inflected Affective Lexicon (MAL, Vassallo et al. 2019) and the Weighted Morphologically-inflected Affective Lexicon (WMAL, Vassallo et al. 2020). Sentix comprises forty-one thousand headwords in their basic form, omitting all the possible inflected forms, therefore lemmatisation is an important step that needs to be taken. MAL is an extended version of Sentix in which lexical entries associated with polarity ratings are given in their inflected forms, thereby dispensing with the lemmatisation process.

[4] NodeXL: Network Overview, Discovery and Exploration for Excel, https://nodexl.com/.
[5] Socialgrabber, the Twitter Extraction tool, https://www.socialgrabber.net.
[6] https://github.com/edgresearch/dataset-sentiment-ita-mattarella2018.

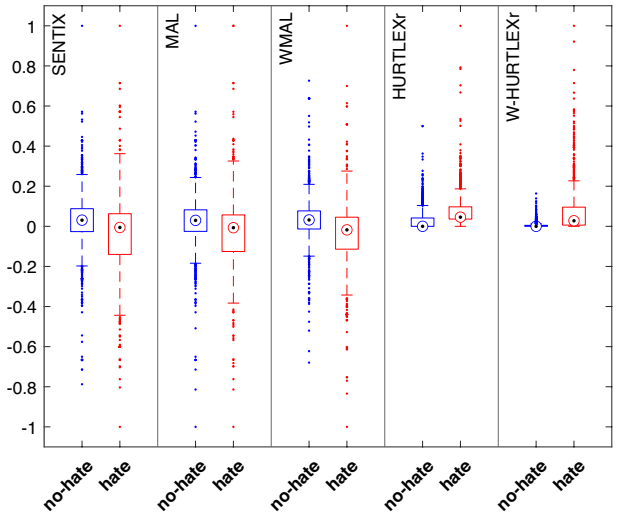

Fig. 7 Boxplot of the score distributions for Mattarella's corpus obtained via Sentiment Analysis (*SENTIX, MAL, WMAL*) and using the revised HurtLex lexicon (*HURTLEXr, WHURTLEXr*). The textual documents are classified according to the presence of hate speech

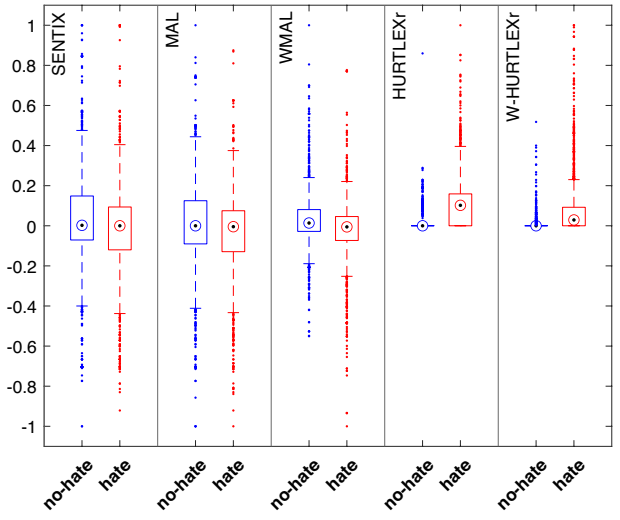

Fig. 8 Boxplot of the score distributions for the Evalita corpus obtained via Sentiment Analysis (*SENTIX, MAL, WMAL*) and using the revised HurtLex lexicon (*HURTLEXr, WHURTLEXr*). The textual documents are classified according to the presence of hate speech

WMAL considers ratings weighted in accordance with the frequency of words' appearance in TWITA,[7] a large scale corpus covering over 500 million tweets (Basile et al. 2018).

The lexicon-based Sentiment Analysis was performed using functions taken from the SentixR R package.[8] We developed an R script for the lexicon-based offensiveness detection using the Tidytext R package (Silge and Robinson 2016).

Figures 7, 8 and 9 represent the score distributions for the three corpora distinct for no-hate and hate comments. More specifically, they show the sentiment scores obtained exploiting the three selected affective dictionaries (i.e. Sentix, MAL and Weighted MAL)

[7] http://twita.di.unito.it/.

[8] https://github.com/valeriobasile/sentixR.

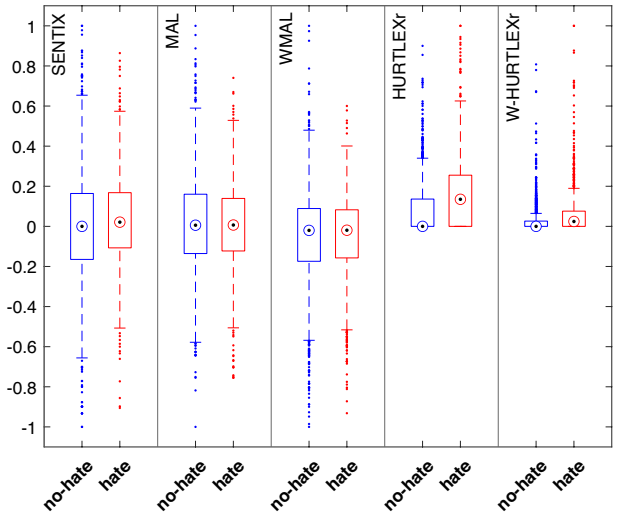

Fig. 9 Boxplot of the score distributions for the migrants corpus obtained via Sentiment Analysis (*SENTIX, MAL, WMAL*) and using the revised HurtLex lexicon (*HURTLEXr, WHURTLEXr*). The textual documents are classified according to the presence of hate speech

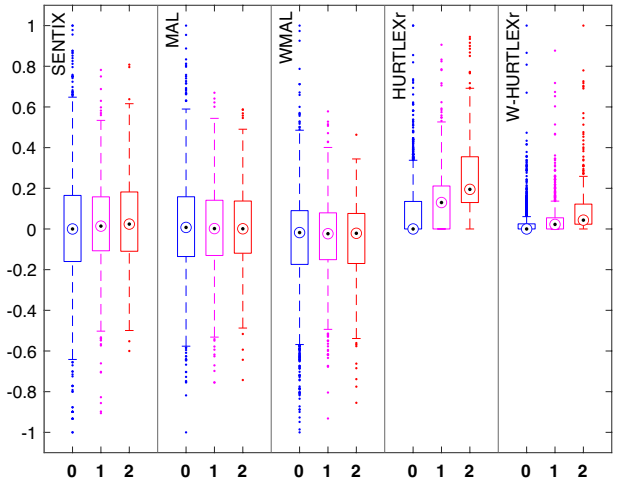

Fig. 10 Boxplot of the score distributions for the migrants corpus obtained via Sentiment Analysis (*SENTIX, MAL, WMAL*) and using the revised HurtLex lexicon (*HURTLEXr, WHURTLEXr*). The textual documents are classified according to the level of intensity of offensiveness: *0 no offensiveness, 1 weak offensiveness, 2 strong offensiveness*

and the scores computed using the offensiveness weights of the Revised HurtLex lexicon. The weighted offensive score is calculated by weighting the offensiveness score by the proportion of abusive words on each comment's total number of words. For comparison, the sentiment scores are scaled between −1 and 1, while the offensive scores, being positive, are scaled in the range [0, 1].

For all the corpora, it is evident how comments not classified as hate speech tend to have positive sentiment scores for all the affective lexicons, while offensiveness scores are near to zero specially if a weighting scheme is used. In contrast, when comments are classified as hate speech, sentence sentiment score tends to decrease and sentence offensiveness score tends to increase.

Figure 10 represents the scores according to the annotation based on the level of offensiveness for the migrants corpus. Comments that do not contain offensiveness show offensiveness scores near to zero, offensiveness scores tend to increase as we go from a weak

Table 1 Results of the fivefold cross-validation experiment

Test set	Model	Accuracy	Precision	Recall	F1-score
Mattarella	BERT	.730	.736	.512	.575
	BERT + HurtLexr	**.737**	.754	.453	.556
EVALITA	BERT	.706	.690	.798	.739
	BERT + HurtLexr	**.736**	.720	.816	.763
IHSC	BERT	.698	.602	.173	.262
	BERT + HurtLexr	**.704**	.630	.192	.288

The bestaccuracy for each test set is highlighted in bold face

level to a strong offensiveness level. On the other hand sentiment scores do not allow to distinguish the different levels of offensiveness.

4.3 Hate speech detection

In order to test the effectiveness of Revised HurtLex on a downstream task, we performed a supervised classification experiment, employing a state-of-the-art based on BERT (Devlin et al. 2019). In particular, we used the pre-trained model `bert-base-italian-cased` from the Huggingface repository,[9] which is trained on a large corpus of Italian text. The experiment is a fivefold cross validation of a binary classification task, where the input text is an instance of one of the three datasets analyzed in this paper, and the label is either *hate speech* or *not hate speech*. As a baseline, we fine-tune BERT on 80% of the data and predict the labels of the remaining, unseen 20%, repeating this process for each of the ten training/test splits. We use a learning rate of $10^{-}5$, a batch size of 4, and a fixed number of epochs of 3.

For testing the impact of the resource introduced in this article, we implemented a variation of the BERT model inspired by HurtBERT (Koufakou et al. 2020). We encoded the words in Revised HurtLex as 17-dimensional vectors, each corresponding to a HurtLexr category, where the scores for each word belonging to a category of HurtLex are the values of the corresponding dimension in the embedding. The HurtLex embeddings are fed into a 8-unit Long Short-term Memory (LSTM) and the output of this additional branch of the network is concatenated to the BERT embeddings, followed by a 0.5 drop-out layer and a single-neuron output layer with sigmoid activation.

The results of the fivefold cross validation are shown in Table 1 in terms of accuracy, precision, recall, and F1-score, averaged over the fivefolds.

It is clear from the results how the inclusion of knowledge from the Revised HurtLex improves the classification performance on all three datasets, however impacting precision and recall differently depending on the test data. Interestingly, while the inclusion of the Revised HurtLex always helps with the accuracy, the performance in terms of F1-score is affected negatively in the case of the Mattarella corpus. In this corpus, the topical focus is much narrower than in the other two corpora. We hypothesize that this contributes to the misalignment between the performance deltas measured in

[9] https://huggingface.co/dbmdz/bert-base-italian-cased.

precision vs recall. With the injection of knowledge from the Revised HurtLex, the recall decreases, possibly because expressions of hate in the Mattarella corpus are more tied to specific people and events and less to typical derogatory words for which the lexicon helps. The extra knowledge still boosts the precision of the classification, although not enough to compensate in terms of F1.

5 Conclusions

The original contribution that our work has to offer is the creation of the Revised HurtLex lexicon, a novel resource in which each headword is annotated with a grading of its offensiveness level derived by exploiting a 2-PNO IRT model. This model has proved to be useful also for evaluating raters' consistency agreement given a rating scale manual annotation scheme. We decided to add a grading to each entry because it could be useful for tracking down offensive and hateful expressions. In fact, the findings in the application to three different textual corpora have shown how the sentence offensiveness level tends to increase when we are in the presence of hate speech or uncivil contents.

The investigation of uncivil online contents based on lexicons has the clear advantage of relating to a large number of derogatory terms and swear words that can be easily detected. Moreover, the lexicon-based approach is fully unsupervised, and therefore easy to apply to other domains and data without the need to manually curate large and expensive training sets for machine learning. On the other hand, the disadvantage of using lexicons is that swear words are used in everyday speech sometimes without offensive intent, therefore their detection may lead to false-positive results in classification. Therefore, it is advisable to use the Revised HurtLex lexicon in conjunction with other techniques. Most studies exploit supervised approaches to distinguish between contents containing or not-containing hate. Recently, hybrid techniques started to show their potential, e.g., by including features extracted from the original HurtLex into a supervised classifier (Koufakou et al. 2020). In this context, the gradings assigned to each word of the Revised HurtLex lexicon could be used to derive even more informative synthetic features to be included in various supervised approaches.

Future research will address the addition of the target topics to the lexicon entries and the evaluation of their performance for hate speech detection. In fact as shown by Chiril et al. (2021), who exploited the categories already present in HurtLex as selected features to train several classifiers, this can help in detecting both the topics (racism, xenophobia, sexism, misogyny) and the targets (gender, ethnicity) of abusive speech.

Author contributions All authors contributed to the study conception and design. Material preparation and data collection were performed by AT, data analysis was performed by all authors. The first draft of the manuscript was written by LF and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Università degli Studi G. D'Annunzio Chieti Pescara within the CRUI-CARE Agreement. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**(422), 669–679 (1993). https://doi.org/10.1080/01621459.1993.10476321

Almatarneh, S., Gamallo, P.: A lexicon based method to search for extreme opinions. PLoS ONE **13**(5), 1–19 (2018). https://doi.org/10.1371/journal.pone.0197816

Baldwin, P., Bernstein, J., Wainer, H.: Hip psychometrics. Stat. Med. **28**(17), 2277–2292 (2009). https://doi.org/10.1002/sim.3616

Basile, V., Nissim, M.: Sentiment analysis on Italian tweets. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 100–107. Association for Computational Linguistic, Atlanta, GA (2013). https://aclanthology.org/W13-1614

Basile, V., Lai, M., Sanguinetti, M.: Long-term social media data collection at the University of Turin. In: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, 10–12 Dec 2018 (2018). http://ceur-ws.org/Vol-2253/paper48.pdf

Bassignana, E., Basile, V., Patti, V.: Hurtlex: a multilingual lexicon of words to hurt. In: 5th Italian conference on computational linguistics, CLiC-it 2018, CEUR-WS, pp. 1–6 (2018). http://ceur-ws.org/Vol-2253/paper49.pdf

BeersFägersten, K.: A sociolinguistic analysis of swearword offensiveness. Saarl. Work. Pap. Linguist. **1**, 14–37 (2007)

Bosco, C., Dell'Orletta, F., Poletto, F., et al.: Overview of the EVALITA 2018 Hate Speech Detection Task. In: Caselli, T., Novielli, N., Patti, V., et al (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, 12–13 Dec 2018, vol. 2263. CEUR-WS.org (2018). http://ceur-ws.org/Vol-2263/paper010.pdf

Buechel, S., Hahn, U.: Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2892–2904. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018). https://aclanthology.org/C18-1245

Chiril, P., Pamungkas, E., Benamara, F., et al.: Emotionally informed hate speech detection: a multi-target perspective. Cogn. Comput. **1**, 31 (2021). https://doi.org/10.1007/s12559-021-09862-5

Coe, K., Kenski, K., Rains, S.A.: Online and uncivil? Patterns and determinants of incivility in newspaper website comments. J. Commun. **64**(4), 658–679 (2014). https://doi.org/10.1111/jcom.12104

Cohen-Almagor, R.: Fighting hate and bigotry on the internet. Policy Internet **3**(3), 1–26 (2011). https://doi.org/10.2202/1944-2866.1059

Cowles, M.: Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. Stat. Comput. **6**(2), 101–111 (1996). https://doi.org/10.1007/BF00162520

de Ayala, R.J.: The Theory and Practice of Item Response Theory. The Guilford Press, New York (2009). https://doi.org/10.1111/j.1745-3984.2010.00124.x

De Mauro, T.: Le parole per ferire. Internazionale. 27 settembre 2016. Compiled for the "Joe Cox" Committee on intolerance, xenophobia, racism and hate phenomena, of the Italian Chamber of Deputies (2016)

Devlin, J., Chang, M., Lee, K., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), Minneapolis, MN (2019). https://doi.org/10.18653/v1/N19-1423

Faloppa, F.: # Odio: Manuale di resistenza alla violenza delle parole. UTET, Turin (2020)

Fersini, E., Nozza, D., Rosso, P.: AMI @ EVALITA2020: automatic misogyny identification. In: Basile, V., D. C, Di Maro, M., et al. (eds) Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, 17 Dec 2020, CEUR Workshop Proceedings, vol. 2765. CEUR-WS.org (2020). http://ceur-ws.org/Vol-2765/paper161.pdf

Frischlich, L., Schatto-Eckrodt, T., Boberg, S., et al.: Roots of incivility: how personality, media use, and online experiences shape uncivil participation. Media Commun. **9**(1), 195–208 (2021). https://doi.org/10.17645/mac.v9i1.3360

Gitari, N., Zuping, Z., Damien, H., et al.: A lexicon-based approach for hate speech detection. Int. J. Multimed. Ubiquitous Eng. **10**, 215–230 (2015). https://doi.org/10.14257/ijmue.2015.10.4.21

Kiritchenko, S., Mohammad, S.: Best–Worst scaling more reliable than rating scales: a case study on sentiment intensity annotation. In: ACL 2017—55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 465–470 (2017). https://doi.org/10.18653/v1/P17-2074

Koufakou, A., Pamungkas, E.W., Basile, V., et al.: Hurtbert: Incorporating lexical features with BERT for the detection of abusive language. In: Akiwowo, S., Vidgen, B., Prabhakaran, V., et al (eds) Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, 20 Nov 2020, pp. 34–43. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.alw-1.5

Maisto, A., Pelosi, S., Vietri, S., et al.: Mining offensive language on social media. Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017, pp. 252–256 (2017). https://doi.org/10.4000/books.aaccademia.2441

O'Sullivan, P.B., Flanagin, A.J.: Reconceptualizing 'flaming' and other problematic messages. New Media Soc. **5**(1), 69–94 (2003). https://doi.org/10.1177/1461444803005001908

Poletto, F., Basile, V., Bosco, C., et al.: Annotating hate speech: three schemes at comparison. In: Bernardi, F., Navigli, R., Semeraro, G. (eds) Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, 13–15 Nov 2019, vol. 2481. CEUR-WS.org (2019). http://ceur-ws.org/Vol-2481/paper56.pdf

Poletto, F., Basile, V., Sanguinetti, M., et al.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Eval. **55**, 477–523 (2021). https://doi.org/10.1007/s10579-020-09502-8

Razavi, A.H., Inkpen, D., Uritsky, S., et al.: Offensive language detection using multi-level classification. In: Farzindar, A., Kešelj, V. (eds.) Advances in Artificial Intelligence Canadian AI 2010. Lecture Notes in Computer Science, pp. 16–27. Springer, Berlin (2010). https://doi.org/10.1007/978-3-642-13059-5_5

Sabatini, F., Sarracino, F.: Online social networks and trust. Soc. Indic. Res. **142**(1), 229–260 (2019). https://doi.org/10.1007/s11205-018-1887-2m

Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychometrika **34**(17), 1–97 (1969). https://doi.org/10.1007/BF03372160

Sanguinetti, M., Poletto, F., Bosco, C., et al.: An Italian Twitter corpus of hate speech against immigrants. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018). https://aclanthology.org/L18-1443

Schmidt, A., Wiegand, M.: 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. Association for Computational Linguistics, Valencia, Spain (2017). https://doi.org/10.18653/v1/W17-1101

Silge, J., Robinson, D.: tidytext: text mining and analysis using tidy data principles in R. J. Open Source Softw. **1**(3), 37 (2016). https://doi.org/10.21105/joss.00037

Skrondal, A., Rabe-Hesketh, S.: Structural equation modeling: categorical variables. In: Everitt, B., Howell, D. (eds.) Encyclopedia of Statistics in Behavioral Science. Wiley, London (2005). https://doi.org/10.1002/0470013192.bsa596

Taboada, M., Brooke, J., Tofiloski, M., et al.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011). https://doi.org/10.1162/COLI_a_00049

Vargas, F.A., Carvalho, I., Rodrigues de G'oes, F.: Identifying offensive expressions of opinion in context (2021). arXiv:2104.12227

Vassallo, M., Gabrieli, G., Basile, V., et al.: The tenuousness of lemmatization in lexicon-based sentiment analysis. In: Bernardi, R., R. N, Semeraro, G. (eds) Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, 13–15 Nov 2019, vol. 2481. CEUR-WS.org (2019). http://ceur-ws.org/Vol-2481/paper74.pdf

Vassallo, M., Gabrieli, G., Basile, V., et al.: Polarity imbalance in lexicon-based sentiment analysis. In: Monti, J., Dell'Orletta, F., Tamburini, F. (eds) Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, 1–3 March 2021, vol. 2769. CEUR-WS.org (2020). http://ceur-ws.org/Vol-2769/paper_36.pdf

Warriner, A., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 English lemmas. Behav. Res. Methods **45**(4), 1191–1207 (2013). https://doi.org/10.3758/s13428-012-0314-x

Wiegand, M., Ruppenhofer, J., Schmidt, A., et al.: 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, vol. 1 (Long Papers), pp. 1046–1056. Association for Computational Linguistics, New Orleans, LO (2018). https://doi.org/10.18653/v1/N18-1095

Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pp. 347–354. Association for Computational Linguistics (2005). https://doi.org/10.3115/1220575.1220619

Zhao, J.L., Li, M.Z., Yao, J., et al.: The development of the Chinese sentiment lexicon for internet. Front. Psychol. (2019). https://doi.org/10.3389/fpsyg.2019.02473