Language Resources and Evaluation (2024) 58:1115–1155
https://doi.org/10.1007/s10579-023-09711-x

ORIGINAL PAPER

NewsCom-TOX: a corpus of comments on news articles annotated for toxicity in Spanish

Mariona Taulé[1] · Montserrat Nofre[1] · Víctor Bargiela[1] · Xavier Bonet[1]

Accepted: 20 November 2023 / Published online: 17 January 2024

Abstract

In this article, we present the NewsCom-TOX corpus, a new corpus manually annotated for toxicity in Spanish. NewsCom-TOX consists of 4359 comments in Spanish posted in response to 21 news articles on social media related to immigration, in order to analyse and identify messages with racial and xenophobic content. This corpus is multi-level annotated with different binary linguistic categories -stance, target, stereotype, sarcasm, mockery, insult, improper language, aggressiveness and intolerance- taking into account not only the information conveyed in each comment, but also the whole discourse thread in which the comment occurs, as well as the information conveyed in the news article, including their images. These categories allow us to identify the presence of toxicity and its intensity, that is, the level of toxicity of each comment. All this information is available for research purposes upon request. Here we describe the NewsCom-TOX corpus, the annotation tagset used, the criteria applied and the annotation process carried out, including the inter-annotator agreement tests conducted. A quantitative analysis of the results obtained is also provided. NewsCom-TOX is a linguistic resource that will be valuable for both linguistic and computational research in Spanish in NLP tasks for the detection of toxic information.

Keywords Toxic language · Subjectivity · Corpus annotation

✉ Mariona Taulé
mtaule@ub.edu

Montserrat Nofre
montsenofre@ub.edu

Víctor Bargiela
vbargiela@ub.edu

Xavier Bonet
xavierbonetcasals@ub.edu

[1] CLiC, Centre de Llenguatge i Computació, University of Barcelona, 08007 Barcelona, Spain

1 Introduction

The automatic detection of toxic language on social media and news websites is a task that has attracted growing interest in the Natural Language Processing (NLP) field in the last few years. Deciding whether the content of a message (a comment or tweet, for instance) is toxic or not is certainly a complex task to address both for humans and especially for automatic systems, mainly due to the nature of toxicity. The interpretation of the message as toxic or not will be determined partially but inevitably by subjectivity, i.e. it is influenced by the beliefs, political ideas and interests of the reader, by how we understand the world, and even conditioned by how we feel at the moment we read that message. In addition, determining whether the message is toxic or not also involves dealing with pragmatics; we have to take into account real-world knowledge to be able to interpret both the explicit (literal) and implicit (inferred) meaning of the message. Finally, we must not overlook the intrinsic ambiguity of language: the same message can be interpreted differently if the user aims to be intentionally ambiguous or if the text does not convey adequately the content because of an incorrect use of punctuation and grammatical errors, which are easy to find, for instance, in comments and tweets posted on social media. Another point to consider is the fact that most of the techniques used for the detection of toxic language rely on machine learning models trained on annotated corpora (Davidson et al., 2017; Schmidt & Wiegand, 2017a; Waseem et al., 2017; Fortuna & Nunes, 2018; Salminen et al., 2020). These models require a large amount of data for learning, but what is most important is to have high-quality corpora annotated for toxicity. The quality of the annotation, therefore, is essential in order to guarantee the quality of these models (Poletto et al., 2021; Vidgen & Derczynski, 2021; Fortuna et al., 2020).

The objective and main contribution of this article is to present NewsCom-TOX, the first corpus of comments in response to news articles posted on social media which have been annotated for toxicity in Spanish. To do so, we must first define the type of contents that are considered toxic and the categories that characterize a toxic message. We then present the guidelines for the annotation of toxicity in which the tagset used, based on the categories identified, and the criteria for assigning the annotation labels are described, as well as the annotation process carried out, including the inter-annotator agreement tests conducted. The main objective is to obtain a high-quality, annotated corpus and use as much contextual information as possible in its process of annotation in order to minimize the subjectivity of the annotators and increase the consistency of the final annotation. We are especially interested in the characterization of toxic language, in the categories that allow us to identify toxic contents and in the creation of the NewsCom-TOX corpus for analysing toxicity from a linguistic perspective and for training and evaluating automatic systems for detecting this type of toxic contents.

The NewsCom-TOX corpus is released under a Creative Commons Attribution ShareAlike 4.0 International license (CC BY-SA 4.0 License) and available for research purposes upon request. The corpus consists of 4359 annotated comments and the whole conversational thread in which each comment occurs, as well as the 21 published news articles to which the comments respond, including their corresponding images. In addition to the gold standard, the pre-aggregated annotation, that is, the individual annotations of the three annotators involved can also be provided if required. We provide these pre-aggregated versions of the corpus in order to be used for applying learning with disagreements approaches (Uma et al., 2021), and "to model the different perspectives that annotators may adopt towards certain highly subjective phenomena" (Akhtar et al., 2021; Abercrombie et al., 2022)

This article is structured as follows: In Sect. 2, we present a brief description of the Spanish language annotated corpora which contain toxic content. In Sect. 3, we first define what we mean by toxic language and then the type of contents that are considered toxic and the categories that characterize these toxic contents. We describe the NewsCom-TOX corpus in Sect. 4. Section 5 is devoted to the annotation guidelines (5.1) and the annotation process, including the inter-annotator agreement tests performed (5.2), and the description of the hypotheses on which the selection of categories to be annotated are based (5.3). In Sect. 6, we describe the results of the annotation. Finally, our conclusions and future work are set out in Sect. 7.

2 Related work: Spanish language corpora containing toxic content

This section provides a brief description of the Spanish annotated corpora currently available for toxic, abusive, aggressive and hate speech content. We focus on the corpora in Spanish because Vidgen and Derczynski (2021)[1] and Poletto et al. (2021) already offer good overviews of corpora annotated with this kind of information for other languages, as well as three of the corpora in Spanish that we present below. If we consider the data referenced in the above overviews, almost 50% of datasets are for English and a little more than 50% are gathered from Twitter, which is the most used social media source.

Table 1 summarizes the ten available Spanish corpora that have been annotated with toxic or abusive content. For the sake of completeness and a better comparison of corpora, we have also added the NewsCom-TOX corpus to this table, although we will describe this resource in more detail in the following sections. Most of the corpora described are used as datasets in different shared tasks, mainly at IberEval_2018,[2] IberLEF_2019,[3] 2020,[4] 2021[5] and 2022,[6] and SemEval_2019.[7].

[1] Vidgen and Derczynski (2021) have created a new open website that lists publicly available abusive language datasets: http://hatespeechdata.com/.

[2] https://sites.google.com/view/ibereval-2018/workshop

[3] https://sites.google.com/view/iberlef-2019

[4] https://sites.google.com/view/iberlef2020/home

[5] https://sites.google.com/view/iberlef2021/home

[6] https://sites.google.com/view/iberlef2022/home

[7] https://alt.qcri.org/semeval2019/

Table 1 Spanish corpora annotated with toxic content

Dataset and Reference	Source	Size	% toxic content	Task workshop	Topic target	Annotation strategy	Annotators
AMI 2018 (Fersini et al., 2018)	Twitter	4138	49.8%	Misogyny IberEval-2018	Women	Multi-level	Crowdsourcing +3 experts
MEX-A3T 2018, 2019 (Alvarez-Carmona et al., 2018)	Twitter	11,856	29.6%	Aggressiveness IberEval-2018 IberEval-2019	Generic	Binary	2 experts
MEX-A3T 2020 (Aragón et al., 2020)	Twitter	10,475	28.7%	Aggressiveness IberEval-2020	Women	Binary	2 experts
HateEval 2019 (Basile et al., 2019)	Twitter	6600	41.5%	Hate Speech SemEval-2019	Women immigration	Multi-level	Crowdsourcing 2 experts
HaterNet 2019 (Pereira-Kohatsu et al., 2019)	Twitter	6000	26%	Hate speech	Non-specified	Binary	4 experts
EXIST 2021 (Rodríguez-Sánchez et al., 2021)	Twitter and Gab	5701	50.23%	Sexism IberLEF-2021	Women	Multi-class	Crowdsourcing +5 experts
EXIST 2022 (Rodríguez-Sánchez et al., 2022)	Twitter and Gab	6226	50.08%	Sexism IberLEF-2022	Women	Multi-class	6 experts
OffendES (Plaza-del-Arco et al., 2021)	Twitter, YouTube Instagram	30,416	12.79%	Offensiveness IberLEF-2021	Generic	Multi-level	3–10 experts
OffendMEX (Plaza-del-Arco et al., 2021)	Twitter	7319	27.62%	Offensiveness IberLEF-2021	Generic	Multi-class/Binary	3 experts
NewsCom-TOX (Our proposal)	News websites	4359	31.87%	Toxicity IberLEF-2021	Immigration	Multi-level	4 experts

As can be seen, Twitter is the main source of information for all corpora except for the OffendES (Plaza-delArco et al., 2021) and EXISTS 2021 and 2022 (Rodríguez-Sánchez et al., 2021, 2022) corpora, which also include posts from other social networks in addition to Twitter posts, concretely from YouTube and Instagram in OffendES and from Gab in EXIST 2021 and 2022, although in the latters only 490 out of 5701 posts (EXIST 2021) and 6226 posts (EXIST 2022) included in the datasets are retrieved from Gab. The distribution of posts is not indicated in OffendES. The NewsCom-TOX corpus extracts all of its data from news websites. The size of these datasets ranges from 30,416 tweets (the OffendES corpus) to 4138 tweets (the AMI corpus (Fersini et al., 2018)) and, on average, 38.80% of the contents of these messages is toxic or abusive. EXIST is the corpus with the most balanced data, along with AMI and HateEval (Basile et al., 2019), whereas the other datasets are more in line with the percentages found in social media. The topics are focussed mainly on xenophobic and misogynistic contents, that is, messages in which the toxic or abusive contents target immigration (HateEval and NewsCom-TOX) and women (AMI, MEX-A3T_2020 (Aragón et al., 2020), HateEval and EXIST) respectively. The MEX-A3T_2018 dataset (Álvarez-Carmona et al., 2018) includes tweets related to politics, sexism, homophobia and discrimination (the authors do not specify which kind of discrimination). The HaterNet corpus (Pereira-Kohatsu et al., 2019) does not explicitly specify which type of topics are addressed.

With respect to the annotation included, the different versions of the MEX-A3T datasets and HaterNet corpus are annotated using a binary scheme, in which the messages are tagged with two mutually-exclusive values for identifying the presence or absence of aggressiveness or hate speech. In contrast, the remaining corpora are annotated with a non-binary scheme and for several phenomena at the same time. In the MEX-A3T (2018, 2019 and 2020) corpora, both offensive messages, which include rude, derogatory, pejorative and profanity terms, and more aggressive or hateful messages, which attack an individual or a group or incite violence (without specifying whether it is a minority, protected or stereotyped group), are labelled as 'aggressive'. The annotation criteria applied in MEX-A3T 2020 were revised following the proposal of Díaz-Torres et al. (2020), who proposed more specific criteria for distinguishing between 'aggressive', 'offensive' and 'vulgar' language. The main difference between 'offensive' and 'aggressive' language is that the latter "seeks to harm or hurt a group or individual by referring to or inciting violence" (Díaz-Torres et al., 2020: p. 134). The difference between 'offensive' and 'vulgar' language is that the message is 'offensive' when there is an intention to offend, to insult or to hurt, whereas a 'vulgar' message can be considered an informal message without the intention to offend and may not refer to an individual or group. Aggressive and offensive messages, unlike vulgar messages, require a target in order to be annotated as such and may also contain coarse or profanity language. This type of messages are labelled 'aggressive', whereas vulgar tweets are labelled 'non-aggressive' in the MEX-A3T corpora. The same annotation scheme was followed in the OffendMEX corpus, distinguishing between 'aggressive' and 'vulgar' language, but aggressive and offensive tweets were further grouped and labelled 'offensive' and tweets containing vulgar language 'non-offensive'. In the OffendES corpus, however, messages were labelled 'offensive' and 'non-offensive' according to five different classes distinguishing the type of target to whom the offense was directed -person ('offensive-target-person'), stereotyped or minority groups ('offensive-target-group'), neither of both ('offensive-target-organization')-, while distinguishing non-offensive posts that include rude words, blasphemes or swearwords (vulgar language but non-offensive, 'non-offensive-expletive-language') from those posts in which neither offensive nor vulgar language is used ('non-offensive'). In a nutshell, although the MEX-A3T (2018, 2019, 2020) and Offend (2021) corpora distinguish between aggressive and offensive messages, when annotating these types of messages are grouped under the same aggressive label in the former corpora and the offensive label in the latter. And all of them are considered expressions of abusive language. Aggressiveness, however, is considered a subtype of hate speech in the HateEval corpus, in which those hateful messages that include discriminatory attitudes, potential threats, overt hostility or violent actions are later labelled 'aggressive' or 'non-aggressive'. In contrast, 'offensive' messages are focused on "the potentially hurtful effect of the tweet content on a given target (...) associated with typical human flaws" (Sanguinetti et al., 2018a: p. 2800). In fact, offensive language targeting immigrants, when it includes, for instance, blasphemy or rude language but not hateful content, is not labelled as a hateful message in HateEval. In this corpus, "a message that spreads, incites, promotes or justifies the target, or a message that aims at dehumanizing, hurting or intimidating the target", in which the target are immigrants or women, is labelled a hateful message (Basile et al., 2019). Messages targeting an individual but considering them a member of a stereotyped category rather than not for their individual characteristics are also labelled hateful. In the same way, in the HateEval dataset, the target of the message is annotated as 'individual' or 'generic'.[8]. The authors consider essential, therefore, the presence of a stereotyped target on the basis of characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion or other in order for a message to be tagged hateful. In the same vein, the HaterNet corpus is annotated with hate speech specifically focusing on those expressions that constitute a criminal offence and/or "expressions that are not criminally punishable, but may justify a civil suit or administrative sanctions" (Pereira-Kohatsu et al., 2019: p. 2), based on Rabat Plan of Action of the United Nations,[9]. that is, hate crimes targeting minority or protected groups.

The AMI MEX-A3T (2020) and EXIST (2021, 2022) corpora are, like HateEval, examples of annotated corpora that focus on a special type of hate speech targeting a specific social group, in this case, misogynistic or sexist messages, in which the hateful messages target women. In AMI, misogynistic messages are defined as hate or prejudice against women, including social exclusion, discrimination, hostility, threats of violence and sexual objectification. Misogynistic tweets are tagged into five different classes considering the type of misogynistic behaviour, namely 'stereotype and objectification', 'dominance', 'derailing', 'sexual harassment and threats of violence', and 'discredit'. In addition, the type of target to which the misogynistic

[8] https://github.com/msang/hateval/blob/%20master/annotation_guidelines.md

[9] https://www.ohchr.org/en/freedom-of-expression

tweet is addressed is also tagged in terms of two values: ‘active (individual)’, when the message targets a specific woman, ‘passive (generic)’, when it targets a group of women (Fersini et al., 2018: pp. 215–216). Similarly, sexist messages are annotated in the EXIST corpora by considering different types of sexist content or behaviour, such as ideological issues, stereotyping, sexual violence and objectification, which are categorized under the following five labels: ‘ideological and inequality’, ‘stereotyping and dominance’, ‘objectification’, ‘sexual violence’ and ‘misogyny and non-sexual violence’. In this corpus, implicit (indirect) expressions of sexism, including any type of sexist expression, such as “descriptive or reported assertions where the sexist message is a report or a description of a sexist behavior” are also considered (Rodríguez-Sánchez et al., 2021: pp. 196–197). The HateEval corpus also deals with another particular type of hate speech, xenophobic messages, in which the hateful messages target immigrants.

Considering the three different strategies for the annotation of data proposed by Poletto et al. (2021), the MEX-A3T and HaterNet corpora follow the first strategy of annotation based on a binary scheme [i.e. two mutually exclusive values, for instance, two values for indicating the presence (*yes*) or absence (*no*) of hate speech in HaterNet]. In contrast, the OffendES and OffendMEX corpora follow the second strategy and the AMI, HateEval and EXIST corpora follow the third strategy. The second strategy suits those corpora that apply a non-binary annotation scheme, that is, more than two mutually exclusive values (for instance, to distinguish between *aggressive*, *offensive* or *vulgar* messages in the Offend datasets) or non-exclusive values, accounting either for different shades of a given phenomenon or for several phenomena at the same time [such as *racism, sexism, both, neither* in Waseem and Hovy (2016)]. The third strategy is based on multi-level annotation, with finer-grained schemes accounting for different phenomena. This annotation scheme involves both a number of different traits and a scale of variation (for instance, to distinguish different types or subclasses of misogyny and sexism in AMI and EXIST respectively).

Finally, all the corpora are annotated by experts, except for the AMI, HateEval and EXIST 2021 datasets, in which the annotation is carried out by expert and crowdsourced annotators. The NewsCom-TOX corpus is described in detail in Sects. 4 and 5.

3 What do we mean by toxic language?

The complexity of toxicity is also reflected in the very definition of the phenomenon. Several proposals for a definition have been made, which are materialized in different terms used to refer to it, such as hate speech, offensive, abusive, aggressive or toxic language (Fortuna & Nunes, 2018; Poletto et al., 2021) and (Vidgen & Derczynski, 2021). In fact, these different terms reflect the need to establish the boundaries for considering a content to be toxic or not. However, this diversity of terms may result, in some cases, in ‘fuzzy boundaries’ and an overlapping between the different definitions of these terms.

We consider that a comment is toxic when it attacks, hurts, threatens, insults, offends, denigrates or disqualifies a person or group of people on the basis of characteristics such as race, ethnicity, nationality, political ideology, religion, gender and sexual orientation, among others, regardless of whether the writer intends to be hurtful or offensive (Mall et al., 2020). This attack can be expressed in different ways -explicitly (through insult and inappropriate language) or implicitly (for instance through sarcasm and mockery)- and at different levels of intensity, that is at different levels of toxicity (from impolite and offensive comments to the most aggressive, the latter being those comments that incite hate or even physical violence) (Taulé et al., 2021). It is worth noting that this definition can be applied to toxic spoken, written and signed language (McTavish 2013), although we will focus on the analysis of toxic written language and, concretely, on online toxic language targeting immigrants.

We use toxicity as an umbrella term under which different definitions used in the literature to describe offensive (Zampieri et al., 2020); hateful (Nockleby, 2000; Waseem & Hovy, 2016; Schmidt & Wiegand, 2017b; Davidson et al., 2017) and aggressive (Kumar et al., 2018; Sanguinetti et al., 2018b) language can be included. For instance, those messages annotated as aggressive (Álvarez-Carmona et al., 2018; Aragón et al., 2020; Díaz-Torres et al., 2020), hateful (Basile et al., 2019; Pereira-Kohatsu et al., 2019; Rodríguez-Sánchez et al., 2021) or offensive (Plaza-delArco et al., 2021) in the corpora presented in the previous Sect. 2 would be annotated as toxic messages, but with different levels of toxicity, in our proposal of annotation. Offensive messages containing rude, derogatory, pejorative and profanity terms (Nobata et al., 2016) would probably be classified as mildly toxic, whereas hateful and aggressive messages would probably be annotated as toxic and very toxic respectively, following our classification of toxicity levels (see Subsect. (5.1). Nevertheless, a more detailed comparative study of the annotations carried out on these corpora would be helpful. Abusive language is also used as a generic term to group "hate speech, derogatory language, profanity, toxic comments, racist and sexist statements" (Caselli et al., 2020). Nobata et al. (2016) and Founta et al. (2018) also considered hate speech, derogatory and profanity language to be types of abusive language, while, in Poletto et al. (2021), abusive and toxic language are treated as close synonyms.

Based on these definitions, we can identify a first set of categories that allow us to characterize and, therefore, to identify whether a message is toxic or not, such as the presence of insults, mockery, sarcasm, inappropriate language, intolerance and aggressiveness or the threat of violence, but also the target of the toxic message and the level of toxicity. All of these categories have been considered for the annotation of toxicity. The problem is that most of these categories can in turn be subjective, i.e. the determination of whether a message is sarcastic, contains mockery or expresses intolerance, also relies on the interpretation of the reader, as well as on pragmatic knowledge, which is crucial, for instance, to infer the implicit content. In Sect. 5, we describe these categories, and their corresponding labels, in detail.

4 The creation of NewsCom-TOX corpus

The NewsCom-TOX corpus consists of 4359 comments posted in response to different articles extracted from Spanish online newspapers and discussion forums from August 2017 to September 2020. Therefore, the variant of Spanish used in the comments corresponds to European (or Peninsular) Spanish. The corpus is manually annotated for toxicity. The 21 articles selected are related to immigration, because we are interested in analyzing toxic messages that contain, specifically, racial and xenophobic content. We manually selected articles that could potentially lead to controversy with the aim of finding comments with opposing opinions and examples of toxic language. The selected topic tends to attract more toxic comments than other types of topics (Mall et al., 2020).

The 21 articles selected were published in 12 different online newspapers, and the corresponding comments come from both those posted in the same online newspaper in which the article was published, and from other social news websites, such as Menéame,[10] and discussion forums, such as ForoCoches,[11] in which users can comment on the news posts. We used Menéame and ForoCoches when comments in online newspapers were blocked and when there were too few comments. These articles were manually selected taking into account their controversial subject matter, their potential toxicity, and the number of comments posted (minimum 50 comments). We used a keyword-based approach to search for articles related to immigration. Those keywords were the following: '*inmigración*' (immigration), '*inmigrante*' (immigrant), '*MENA*'[12], '*musulmán*' (muslim), '*negro*' (black), '*patera*'[13] (refugee boat), '*racismo*' (racism), '*racista*' (racist), '*refugiado*' (refugee), '*xenofobia*' (xenophobia) and '*xenófobo*' (xenophobe).[14] Once the articles were selected, we further classified them into three more specific groups according to the topic covered in the news article (i.e. 'migration', 'criminality' and 'society', described below).

The comments were selected in the same order in which they appear in the time thread in the web. The author (pseudonymised) and the date and the time at which the comments were posted are also retrieved, as well as the conversational thread in which the comments were presented.

Table 2 shows the distribution of comments by date of publication, topic, the title of the article[15] selected and the newspaper or discussion forum from which they were obtained. Each file contains the whole comments posted to the corresponding

[10] https://www.meneame.net/

[11] https://www.forocoches.com/

[12] MENA (Menores Extranjeros No Acompañados) is an acronym used to refer to unaccompanied foreign minors under 18 years old, who are in Spain without the care or supervision of an adult.

[13] *Patera* stands for a precarious craft typically used by refugees. We will translate this term as refugee boat.

[14] We also searched for these keywords by their corresponding inflection in plural and, in the case of '*musulmán*', '*refugiado*' and '*xenófobo*', also by their feminine inflection (that is, '*musulmana*', '*refugiada*' and '*xenófoba*').

[15] To save space, we have included the titles translated to English. The original titles in Spanish can be found in Table 11.

Table 2 Distribution of comments per topic in the NewsCom-TOX corpus CR=Crime; MI=Migration; SO=Society

Date	Topic	News article title	Comments source	Comments
20171908	CR	Investigators place Ripoll’s new Imam at the head of the terrorist group	https://www.eldiario.es	38
			https://www.meneame.net	161
20190512	MI	At least 63 deaths after a refugee boat sinks off Mauritania on its way to the Canary Islands	https://www.elpais.com	0
			https://www.forocoches.com	359
20190513	MI	Government asks Church to take in refugees rejected by the Left	https://www.abc.es	191
20190519	SO	Why do victims of male violence can trigger/prompt toxic relationships?	https://www.elpais.com	172
20190716	CR	A ‘head hunter’ in the refugee boat	https://www.elmundo.es	0
			https://www.forocoches.com	320
20190919	MI	The Aragonese government is looking for seven chalets to house unaccompanied migrant minors for 300,000 euros per month	https://www.abc.es	0
			https://www.forocoches.com	132
20200403	MI	150,000 people are needed to work in the countryside	https://www.elmundo.es	0
			https://www.forocoches.com	221
20200424	MI	More than 5,000 euros to escape from Spain by boat	https://www.elmundo.es	162
20200608	SO	David Cantero goes viral for 27 s: his comments on immigrants in Spain	https://www.ecoteuve.eleconomista.es	237
20200618	SO	The young immigrant squatters of Premiá have more than twenty priors for violent robbery	https://www.lavanguardia.com	0
			https://www.meneame.net	65
20200621	SO	Government Subdelegation orders investigation of Javier Negre for a video with street vendor in Galicia	https://www.eldiario.es	86
20200622	MI	Young Maghrebi beaten up after a violent robbery of an old woman	https://www.niusdiario.es	0
			https://www.meneame.net	194
20200626	SO	Lacasa forced to clarify that Conguitos are peanuts and not people	https://www.libremercado.com	0
			https://www.meneame.net	200
20200705	SO	‘The customer of a bar to the police: “I don’t want to share a terrace with these black bitches” ’		132
20200708	MI	Unidas Podemos proposes the regularization of 600,000 undocumented immigrants who were in Spain when the state of alarm was declared	https://www.elmundo.es	328
20200715	CR	Man has finger chopped off with machete after complaining about noise in a park in Madrid	https://elcaso.elnacional.cat	273

Table 2 (continued)

Date	Topic	News article title	Comments source	Comments
20200715b	CR	The TSJ of Extremadura expels from Spain the man who bought his 13-year-old wife for 50€	https://www.extremadura7dias.com	0
			https://www.meneame.net	129
20200725	CR	Two minors strangle a woman and steal her mobile phone in a tunnel in Batán: “They left me unconscious”	https://www.abc.es	256
20200726	SO	Public prosecutor in Málaga demands the closure of Alerta digital for inciting hate	https://www.eldiario.es	0
			https://www.meneame.net	125
202008031	CR	At least four arrested for abusing and assaulting an underage girl at Wifi park	https://www.ultimahora.es	0
			https://www.meneame.net	233
20200901	CR	Trump defends 17-year-old arrested for killing two protesters in Wisconsin: “He fell and was attacked”	https://www.20minutos.es	0
			https://www.mename.net	345
Total		21 articles		4359

news article, and the name of the file indicates the date in which they were posted and the topic of the news article. We distinguish three main topics: (a) **migration**, to indicate that the news article is directly related to events concerning the arrival and integration of migrants or refugees (e.g. the sinking of a refugee boat, refugee hosting, legalization or regularization of immigrants); (b) **criminality**, to indicate that the article is related to crimes in which immigrants are involved (e.g. acts of terrorism, violence or robberies) and (c) **society**, to indicate that the article reflects racist attitudes and behaviors towards immigrant collectives. The number of comments ranges from 65 to 359 comments per article. Most of the comments were posted during the 24–48 h immediately after the news articles were published. In total, 31.87% of the comments are toxic, which is in line with the percentage of toxic messages in other corpora (see Subsect. 6.1).

Regarding the source of information, the online versions of mainstream newspapers, such as ABC, La Vanguardia, El Mundo and El País, tend to filter toxic comments strictly, especially in recent years (in which the filtering has increased), whereas other social news websites and forums, such as Menéame and ForoCoches, apply fewer filters in principle, especially ForoCoches. That explains why 57% of the selected comments were extracted from these websites. In fact, the highest number of very toxic comments come from ForoCoches. Highly toxic comments account for 2.26% of the total comments extracted from this discussion forum, which is higher than the overall average of the corpus (1.81%), as we will see in Subsect. 6.1.

5 Annotating toxicity

Deciding whether the content of a message (such as a comment or tweet) is toxic or not can be determined by different factors, mainly we have to take into account not only the linguistic, but also the extralinguistic context.

The **discursive** content refers to the linguistic content proper, and we need to distinguish between the information conveyed (what is being said) and the way in which it is conveyed, that is, the kind of language that is used to express the content of the message (for instance, improper language, offensive language, rude vocabulary, belittling language, irony, sarcasm and mockery). We also need to take into account that the comment is usually part of a **discourse thread**. This means that a comment may refer to the news directly or to a previous comment and, in the latter case, a conversation or discussion between different users can emerge. Therefore, this discourse or conversational thread must also be taken into account when we annotate the comment, since its interpretation may be conditioned by it. For instance, Pavlopoulos et al. (2020) investigated whether the conversational context conditions annotators' judgements and concluded that these judgements tend to change when the context is taken into account in the annotation of toxicity.

The **extralinguistic context**, however, is related to real-world knowledge. In the case of comments, the extralinguistic context consists of the political, economic, social and cultural events happening at the same time as the publication of the article and the corresponding comments, and makes it possible to interpret them in the most suitable way. The problem is how to access or dispose of the extralinguistic context. This is a challenging task because it depends largely on the world knowledge of the annotators. In order to mitigate the effect of the lack of knowledge of the extralinguistic context on the interpretation of comments, we decided that the annotators should be required to read first the news article which originated the comments annotating them. The news article provides us part of the extralinguistic context: when and where the event took place, the players involved, among others (see Subsect. 5.2). The images that accompany the news articles were also decisive for interpreting the content of the comments, especially for identifying sarcasm and mockery. We also tried to avoid, as much as possible, the bias that may be introduced by any ideological tendency or preference of the annotators (Waseem & Hovy, 2016). To minimize this bias, the ideal team of annotators would consist of people with different beliefs and political views and from diverse backgrounds, genders and ages, although that is not always feasible (see Subsect. 5.2).

In a nutshell, we took into account both the information conveyed in each comment and the discourse thread in which the comment to be annotated occurred, as well as the information conveyed in the news articles and the images accompanying them. The annotators could access all the comments belonging to the same discourse thread, and to the corresponding news articles and images, when they annotated each comment. All this information is available as part of the NewsCom-TOX corpus.

With the aim of reducing subjectivity, or at least inconsistency in the annotation of toxicity and, therefore, also the disagreement between annotators, we proposed first annotating different linguistic categories such as stance, target, stereotype, sarcasm, mockery, insult, improper language, aggressiveness and intolerance, always taking into account the discourse thread in which the comment appears and the information contained in the news articles (including images). These binary categories also allow us to discriminate the level of toxicity of the comments. Furthermore, some of these categories can be correlated, for instance insult and improper language, and these correlations are useful when assigning the level of toxicity. Our general (or starting) hypothesis is that the category types and their combination help to determine the level of toxicity in a more consistent way.

The selection of these categories, which can be split into two main groups -linguistic and contextual categories (subsections 5.1.1 and 5.1.2 respectively)- relies on proposals made in previous works: for instance, target (Waseem et al., 2017; ElSherief et al., 2018; Fersini et al., 2018; Basile et al., 2019), stance (Bosco et al., 2016; Mohammad et al., 2016; Taulé et al., 2017, 2018; Cignarella et al., 2020); stereotype (Allport, 1954; Beukeboom & Burgers, 2019; Sánchez-Junquera et al., 2021); insult (Wulczyn et al., 2017; Dynel, 2021), sarcasm (Farias & Rosso, 2017; Vidgen et al., 2019), level of toxicity [Jigsaw Toxic Comment Classification Challenge 2019[16]; Kolhatkar et al. (2020)]. However, the selection of these categories is also based on a previous analysis of the comments that we annotated and based on the hypotheses, which are postulated in Subsection 5.3.[17] Contextual categories mainly include the discourse or conversational thread in which the

[16] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

[17] We present the hypothesis in Subsection 5.3 in order to have them close to their evaluation described in Subsection 6.2.

comment occurs. We display the whole discourse thread, unlike Pavlopoulos et al. (2020), who only considers the previous comment to the comment being annotated, while we take into account both the previous and subsequent comments.

5.1 Annotation scheme

In the following, the tagset used and the criteria applied for the annotation of comments with toxicity is presented. First, we describe the 13 linguistic categories (which correspond to 15 labels) annotated and then we present the contextual information that we take into account for the annotation of these linguistic categories.

5.1.1 Linguistic categories

5.1.1.1 Stance The Stance category indicates whether a comment is favorable, unfavorable or neutral with regard to a topic of discussion which is usually controversial, and which may or may not be explicitly mentioned in the text message (Mohammad et al., 2016; Lai et al., 2018; Taulé et al., 2017) and (Taulé et al., 2018). For the annotation of Stance, we have two different labels: **Positive_stance** and **Negative_stance**, both with binary values. The combination of these values allows us to distinguish between: (a) **favorable** comments (Positive_stance=yes and Negative_stance=no) (example 1)[18]; (b) **unfavorable** comments (Positive_stance=no and Negative_stance=yes) (example 2); (c) **neutral** comments (Positive_stance=no and Negative_stance=no) (example 3) and (d) comments containing **both stances** (Positive_stance=yes and Negative_stance=yes) (example 4). Comments may be relatively long and, therefore, we can find that one part of the comment is favorable to an idea or argument, but disagrees with another opinion expressed in the news article or in another comment.

1. Se trata de una medida a tener en cuenta, que coincide con las últimas decisiones tomadas por Portugal e Italia.
   'This is a measure to take into account, which coincides with the latest decisions taken by Portugal and Italy.'
   (Positive_stance=yes and Negative_stance=no)
2. No debes inventar que la convivencia es problemática.
   'You should not try to sell the idea that coexistence is problematic.'
   (Positive_stance=no and Negative_stance=yes)
3. Hay un vídeo donde Abascal da los datos de quienes reciben ayudas, y eso que las ayudas solo se conceden a quien tiene papeles, ya me dirás.
   'There is a video where Abascal gives the surname of those who received the subsidies, and given that the subsides can only be granted to those who have papers, what sense does that make.'

[18] All examples were paraphrased in order to ensure the anonymity of the author of the comment and we used * in place of insults, slurs and profanity words. Some of these paraphrases may not reflect the original text as clearly.

(Positive_stance=no and Negative_stance=no)

4. Coincido con el argumento de que el trabajo debe darse primero a los españoles para acabar con las altas tasas de paro. Una vez arreglado esto, los extranjeros ya podrían recibir permiso de trabajo y acceso a la sanidad y la educación públicas mientras tengan trabajo.
   'I agree that work should be given first to Spanish people to put an end to high unemployment rates. Once this is in place, foreigners could get work permits and access to public health and education as long as they are in work.'
   (Positive_stance=yes and Negative_stance=yes)

5.1.1.2 Target The category target also has two labels: **Target_person** and **Target_group**, both of which are binary, for indicating to whom the comment is addressed. The combination of values results in the four following possibilities found in comments: (a) comments targeting a **person** (Target_person=yes and Target_group=no) (example 5); (b) comments targeting a **group** or collective (Target_person=no and Target_group=yes) (example 6); (c) comments targeting both a person and a group (Target_person=yes and Target_group=yes) (example 7), and (d) comments that are not addressed to any person or group in particular (Target_person=no and Target_group=no) (example 8). When the target is a person, that person may be the writer of the news article, the author of another comment or another person who is referred to (example 5).

5. Echenique[19] dice cosas sin sentido; debería regresar a su país!
   'Echenique speaks nonsense; he should go back to his country!
   (Target_person=yes and Target_group=no)
6. Los inmigrantes vendrán a España para contribuir con sus impuestos, ?'qué puede ir mal?
   'Immigrants will come to Spain to contribute with their taxes, what could go wrong?'
   (Target_person=no and Target_group=yes)
7. Se deberían legalizar a los inmigrantes ilegales e ilegalizar a los votantes de Podemos, dirigidos por Echenique e Iglesias.
   'We should legalize the illegal immigrants and illegalize Podemos supporters, led by Echenique and Iglesias.[20]'
   (Target_person=yes and Target_group=yes)
8. En España lo que abunda son las pequeñas empresas, no las multinacionales que ofrecen empleo muy cualificado.
   'In Spain what is abundant are small companies, not multinationals that offer highly skilled employment.'
   (Target_person=no and Target_group=no)

[19] Echenique is a politician in the Spanish left-wing political party Podemos.

[20] Iglesias is a politician in the Spanish left-wing political party Podemos.

5.1.1.3 Stereotype (values=yes/no) Stereotypes are defined as beliefs and ideas widely attributed to a group, by which the individuals in this group are characterized in an undifferentiated and simplified way based on the magnification or exaggeration of an individual characteristic (race, ethnicity, religion, gender, sexual orientation and age, among others) (Allport, 1954; Beukeboom & Burgers, 2019). We annotate a comment as Stereotype=yes when it contains a stereotype, such as example (9), and the absence of stereotypes is annotated as Stereotype=no.

9. No es lógico legalizar a personas que solo quieren vivir de subvenciones. Todo lo que hacen es comer, beber, tener hijos y quejarse de las ayudas.
   'It is illogical to legalize people who only want to live on subsidies. All they do is eat, go out to drink, have children and complain about subsidies.'
   (Stereotype=yes)

5.1.1.4 Sarcasm (values=yes/no) A comment is sarcastic when the content is ironic -that is, when the writer uses words that mean the opposite of what he really wants to say- and when it is accompanied by a harsh, sharp and negative criticism and made in bad faith (Sarcasm=yes) (example 10) (Farias & Rosso, 2017; Vidgen et al., 2019). Ironic comments without intention to cause pain (without a negative load) are not considered toxic and are tagged as Sarcasm=no. We also annotate as sarcastic rhetorical questions accompanied by mockery and sharp criticism, as well as sarcastic jokes and word games. However, genuine jokes with a humoristic intention are annotated as Sarcasm=no.

10. Los votantes de Vox harán los trabajos poco cualificados que actualmente realizan los inmigrantes, lo están deseando.
    'Vox voters are looking forward to doing the low skilled jobs currently done by immigrants.'
    (Sarcasm=yes)

5.1.1.5 Mockery (values=yes/no) This category indicates that the comment ridicules, humiliates or mocks a person or group (Mockery=yes) (example 11).

11. Menudo corte te han metido! Lástima que tu padre no tuviera acceso a los anticonceptivos.
    'What a slap in the face! It's a shame that your father didn't have enough money to access birth control.'
    (Mockery=yes)

5.1.1.6 Insult (values=yes/no) This category indicates that the comment contains one or more explicit insults or slurs with the intention to offend a person or group (example 12). This attribute correlates with the improper language category, which means that if the comment contains insults (Insult=yes) it is also annotated as improper language, but not the other way around (i.e. improper language does not necessarily imply insults or slurs). There are certain words and expressions that we will not consider insults a priori, however when they appear with a clear offensive intentionally they are annotated as Insult=yes (for instance, *gentuza, 'rabble people'*).

12. Lo que dice es una gilipollez, me considero moralmente superior a xenófobos repulsivos, nazis, terroristas y a ti. El buenismo de subnormales perjudica mucho a las personas poco inteligentes.
    'What you are saying is bullshit, I consider myself morally superior to repulsive xenophobes, Nazis, terrorists and you. Easy do-goodism by subnormals does a lot of damage to unintelligent people.'
    (Insult=yes)

5.1.1.7 Improper language (values=yes/no) The category Improper_language indicates that the comment contains language not consider to be proper or that is vulgar and impolite and/or which includes rude words (example 13). However, we may find comments that include insults or improper language used with a humoristic or positive intentionality, such as '*es un coche de puta madre/*it's a fucking great car'. These comments are annotated as Improper_language=yes, they are not considered toxic (Toxicity=no) (example 14).

13. '¿Por qué alguien que en su país no era nadie nos da órdenes y nos toca las pelotas?
    Why can someone who was dirt poor in his own country give us orders and piss us off here?
    (Improper_language=yes, Toxicity=yes)
14. La publicidad de Conguitos[21] y de ColaCao[22] eran jodidamente de mal gusto.
    'Conguitos and Cola-Cao adverts were fucking ugly.'
    (Improper_language=yes, Toxicity=no)

5.1.1.8 Aggressiveness (values=yes/no) A comment contains aggressive language when it expresses violence or a desire to exercise it in a deliberate and conscious or unconscious way, without necessarily including sarcasm, mockery or insults. This aggressiveness can be expressed in a passive manner, by justifying or empathizing with an aggressive action (example 15), or in an active way, by promoting and encouraging or inciting violence (example 16).

15. Aunque la justicia no actúe, se habrá llevado un buen par de hostias.
    'Even if the justice system does not act, he will remember the slaps he got.'
    (Aggressiveness=yes)

[21] Conguitos is a brand of chocolate products, concretely peanuts covered with dark chocolate. The advertising image for this product was a parodied image of black people.

[22] Colacao is a brand of a chocolate drink whose advertising jingle contains a mild racial slur or stereotype.

16. A estos hdp*** como mínimo habría que expulsarlos o quemarlos por cortarle el pulgar.
    'These bastards should at least be extradited or burned alive for cutting off his thumb.[23]'
    (Aggressiveness=yes)

5.1.1.9 Intolerance (values=yes/no) This category indicates the intolerant attitude of the writer of the comment. That means when he or she expresses intransigence or non-acceptance or rejection of the difference of the 'other' or 'others', such as different traditions, customs, beliefs, religions, skin color, sexual orientation or gender, both when they are addressed to an individual (example 17) and a group (example 18).

17. Es urgente que echemos a Echenique de España.
    'It is urgent that we kick Echenique out of Spain.'
    (Intolerance=yes)
18. Nada justifica que puedan seguir viviendo aquí!
    'Nothing justifies that they can continue to live here!'
    (Intolerance=yes)

5.1.1.10 Toxicity (values=yes/no) A comment is toxic when it attacks, denigrates or disqualifies a person or group on the basis of certain characteristics such as race, ethnicity, nationality, religion, gender and sexual orientation, among others. This attack can be expressed in different ways -explicitly (through insult, mockery and inappropriate humor) or implicitly (for instance through sarcasm)- and at different levels of intensity, that is at different levels of toxicity (the most aggressive being those comments that incite hate or even physical violence). We annotate Toxicity=yes when a comment contains Sarcasm, Mockery, Insults, Improper language, Aggressiveness and Intolerance, that is, when one or more of these categories are tagged as 'yes'. Otherwise, we annotate the comment as Toxicity=no.

5.1.1.11 Toxicity level (values=1/2/3) This category indicates the level of toxicity and we therefore only annotate this category if we have previously annotated the comment as toxic (Toxicity=yes). The level of toxicity is determined by the presence and combination of categories that occur in the comment and that have previously determined the toxicity of the comment (i.e. Sarcasm, Mockery, Insults, Improper language, Aggressiveness and Intolerance). Therefore, when we decide the level of toxicity we consider both: (a) the number of toxic categories that appear in the comment -the more negative categories, the higher the level of toxicity tends to be; and especially (b) the presence of specific highly toxic categories, such as Aggressiveness and Insult, especially when more than one insult appears in the comment and the type of insult (see Sect. 6).

[23] The news article reports that a group of Latin-Americans cut off a man's thumb with a machete when he complained about them making noise.

In example (19), *borricos* is not a strong insult and the expression *hay que joderse* ('What a load of bollocks!') is improper language, therefore, we consider the comment to be mildly toxic (Toxicity_level=1).

19. La corrección política nos transforma en verdaderos borricos que no conseguimos darnos cuenta de lo que pasa en los países d alrededor q nos pasan la mano por la cara. Y también aquí, claro. Hay que joderse. Venga, a pagar y dar servicios para los que se beneficiarán· BORRICOS. 'Political correctness turns us into real jackasses unable to see what is happening in neighboring countries that are better/more advanced than us. And of course, here. What a load of bollocks! Let's pay for and provide services for the new beneficiaries... JACKASSES.' (Insult=yes, Improper_language=yes,Toxicity_level=1)

We assign, however, Toxicity_level=2 to example (20), in which the toxicity is implicitly expressed by the combination of sarcasm and mockery.

20. Los de izquierdas dan gracias a que existan mafias que estafen a los que han muerto en el viaje. ?'control? NO.. eso no.. mejor dejar que los 'pobres' lleguen a nuestro país arriesgando sus vidas utilizando a estos delincuentes. GRACIAS PROGRES por los cincuenta y siete muertos que han llegado y los que quedan por venir por haber 'abierto puertas' (que claro, nunca son vuestras). 'The left is grateful for the existence of the mafias that cheat those who die along the way. Control them? NO... let's not do that... let's allow the "poor" to come to Spain risking their lives using these criminals. THANK YOU LEFTISTS for those fifty-seven dead and all the ones that will come because of "open doors" (which are never yours, of course).'
    (Sarcasm=yes, Mockery=yes, Toxicity_level=2)

Comments annotated with Toxicity_level=3 are those with a higher combination of categories, example (21).

21. A estos hdp*** como mínimo habría que expulsarlos o quemarlos por cortarle el pulgar.
    'These bastards should at least be extradited or burned alive for cutting off his thumb.'
    (Insult=yes, Improper_language=yes, Aggressiveness=yes, Intolerance=yes, Toxicity_level=3)

It should be noted that all these labels are binary (value= yes/no) except Toxicity level, which has three possible values (1=mildly toxic, 2=toxic and 3=very toxic). This annotation allows us to establish fine grained criteria for analyzing and better defining what can be considered a comment with toxic language.

5.1.2 Contextual categories

The contextual information mainly includes the discourse or conversational thread in which the comment occurs. This information is very useful for the annotators since it helps them to better interpret and understand the content of the message. The contextual information is retrieved from the news websites and is displayed in the following way.

Comment_id: a number that indicates the chronological order in which the comment was posted in the time thread on the website.
User_id: the pseudoanonymized name of the author of the comment.
Date: the date on which the comment was posted.
Time: the time at which the comment was posted.
Thread_id: indicates the discourse thread in which the comments are displayed. This information is crucial for the annotators, who use it to reconstruct how the different interventions of authors occurred.

It is worth noting that a comment may refer to the news itself or to a previous comment and, in the latter case, a conversation or discussion between different authors can emerge. We indicate the relationship between comments with a number through which the annotators specify:

- Whether the comment refers to the news itself (henceforth primary comment), in which case the number of the thread_id and the number of the comment_id are the same.
- Whether the comment does not refer to the news directly, but to another previous comment (henceforth secondary comment), in which case the number of the thread_id is the same as the comment_id of the comment that it refers to.

Comment_level: a number that indicates whether the comment is primary (value=1) or secondary (value=2). We always annotate secondary comments with value=2, regardless of the degree of nesting of the comment in the discourse thread.

All this information allows annotators to sort and consult the comments taking into account different criteria in order to improve and facilitate the annotation task. For instance, annotators can consult or display comments chronologically, according to the users who have written them or to the discourse or conversational thread, with the aim of making the annotation more accurate. This contextual information also allows annotators to follow the thread of the conversation or discussion that takes place between different authors, and to recover anaphoric references between comments to facilitate the interpretation of their contents. Having access to this information is also interesting for viewing which comments generate longer or more extensive comment threads and may be useful, for instance, for establishing the attributes that generate more responses and reactions.

Annotators therefore always took into account this contextual information, mainly the discourse thread and the comment level in the annotation process to better determine the values of each linguistic category, for instance, to better interpret the stance or whether the comment is sarcastic or contains mockery or stereotypes.

5.2 Annotation process

In this section, we describe the process followed for the annotation of the NewsCom-TOX corpus, in which each comment (the whole comment) was annotated for the 13 linguistic categories applying the criteria presented in the previous subsections (5.1.1 and 5.1.2). Due to the complexity of the task and the high rate of disagreement detected in the training of the annotators, especially in the annotation of the level of toxicity, we decided that each comment would be annotated by three annotators working in parallel. These **annotators** were two women and two men of different ages ranging from 23 to 53 years (23, 37, 51 and 53 years of age), all of whom were white Europeans. Their native language was European Spanish. Two of them were expert annotators specialized in corpus linguistics and the other two were final year students of Linguistics who were specifically trained for the annotation of this task for three months, for four hours per day. The annotators were paid for this annotation task. These students had participated previously in several annotation tasks carried out by the same research group. They had also participated in the elaboration or improvement of the guidelines, which were developed in an iterative way.

The **annotation process** was carried out in the following way. First the annotators read the news article to which the comments refer to. They could read it directly from the same online newspapers and discussion forums in which the article was published by clicking on a link. In this way if the article contained images they could be visualized and taken into account for the annotation as contextual information. Each comment was then annotated by three annotators. First, the binary linguistic categories -Stance, Target, Stereotype, Sarcasm, Mockery, Insult, Improper_language, Aggressiveness and Intolerance- were annotated and, considering these annotations, the Toxicity and Toxicity_level categories were then annotated. The last category that was annotated is the level of toxicity because, the combination of the above-mentioned binary linguistic categories allows us to determine the degree of toxicity (Toxicity_level) in a more consistent way. Some of these linguistic categories are correlated (for instance, if a comment is tagged as Insult=yes then it will also be tagged as Improper_language=yes). Taking this into account, linguistic categories were displayed in the annotation excel file in a specific order to help the annotators to decide the level of toxicity. The annotators were instructed to annotate the linguistic categories in the following order: Stance, Target, Stereotype, Sarcasm, Mockery, Insult, Improper_language, Aggressiveness, Intolerance, Toxicity and Toxicity_level. The Toxicity_level was, therefore, the last category to be annotated. The annotators always took into account the discourse thread when annotating toxicity in comments, which is very useful for interpreting their content, especially when Stance, Sarcasm and Mockery are annotated. Next, an inter-annotator agreement test was conducted, once all the comments associated to the news article had been annotated. All the comments that referred to the same news article were included in one single file. Table 3 shows the results obtained in the inter-annotator agreement test. For each category, the average observed agreement percentage between all the annotators and their corresponding Krippendorff' alpha (Krippendorff, 2004) are shown. Finally, disagreements were discussed weekly by the annotators and a senior researcher until an agreement was reached.

5.2.1 Analysis of disagreements

As we can see in Table 3, the average total agreement obtained for the Toxicity category was 81.84% (0.59 α), whereas it was 73.58% (0.54 α) for the Toxicity_level. The moderate and low results obtained for the Krippendorff's alpha can be explained by the so-called prevalence problem (Eugenio & Glass, 2004) a behaviour of the agreement coefficients that occurs when the annotation data are highly biased in favour of one category. As Artstein and Poesio (2008) explain, if a large amount of data falls into one category, the expected agreement is very high, so to demonstrate high reliability requires even higher observed agreement; this leads to the paradox that annotators can agree on a high proportion of items and yet reliability coefficients are low. The NewsCom-TOX corpus is annotated with binary categories (12 labels out of 13) and only 1389 comments are toxic (31.87%), which penalizes our results when calculating the alpha coefficient. The annotated linguistic categories are less represented and unbalanced (see Tables 6 and 7).

Regarding disagreements (Table 4), whereas 19.36% of the cases showed partial disagreement (only one of the annotators disagreed), only 4.88% of the cases showed total disagreement between all the three annotators. These results highlight the complexity of the task. We should bear in mind that the interpretation of the whole set of categories that we take into account to establish the level of toxicity is highly subjective.

Despite the subjective nature of this task, the observed agreement obtained for the majority of these categories is higher than 81%; the exception being Negative_stance and Toxicity_level. Regarding Negative_stance, disagreement arises mainly from insufficient context and the use of sarcasm.

For instance in example (22), the disagreement between annotators can be explained because the comment could be interpreted as being in favour, in the sense that the writer interprets 'I have also seen the same video as you....' to mean 'I interpret the video in the same way'. But it may also be interpreted that he is against the headline ('A sensationalist headline if ever there was one') because the sentence that mentions the video is sarcastic and, therefore, ends in ellipses (·), meaning that we have seen the same video, but your interpretation does not fit with what I consider to be the reality of the images.

22. Casi le matan de un coscorrón. Tambien vi ese vídeo· Un par de golpes desganados y tres o cuatro empurriones· Titulo sensacionalista como pocos.
'They nearly killed him with a smack on the head. I have also seen the same video as you.... A couple of pushes and three or four weak punches... A sensationalist headline if ever there was one.'

Table 3 Inter-annotator agreement test

Category	Average observed agreement (%)	Krippendorff's alpha
Positive_stance	92.93	0.24
Negative_stance	75.11	0.31
Target_person	87.78	0.53
Target_group	83.88	0.39
Stereotype	90.21	0.44
Sarcasm	84.76	0.37
Mockery	86.70	0.35
Insult	93.72	0.58
Improper_language	93.09	0.60
Aggressiveness	88.51	0.49
Intolerance	81.99	0.27
Toxicity	81.84	0.59
Toxicity_level	73.58	0.54

Another source of disagreement is that the stance of some comments is multiple, because the writer can respond to another writer in a negative way, but at the same time quoting positively the message of another user, while criticizing the news article. In these cases, the stance is not clear enough and leads to disagreements in the annotation.

Regarding the Toxicity_level, the highest level of disagreement occurs when deciding whether a comment is non-toxic or mildly toxic (Toxicity_level=1), followed by when deciding whether a comment is mildly toxic or toxic (Toxicity_level=2). In fact, disagreements on non-toxic or mildly toxic comments account for 54.92% and disagreements on mildly toxic or toxic comments account for 17.48% of the total disagreements in the annotation of toxicity, whereas disagreements on toxic and very toxic comments (Toxicity_level=3) account for 3.56%.

It is also worth noting that disagreement is lower in those comments in which toxicity is more explicit, meaning cases in which categories such as Insult, Mockery, Intolerance and Aggressiveness appear. Moreover, these categories obtain some of the highest percentages of agreement, as shown in Table 3.

We mainly find three sources of disagreement. The most frequent disagreement is related to **subjectivity**. Annotators interpret the comments differently because their do not share the same world knowledge, beliefs, political ideas and interests. For instance, annotators cannot always understand the anaphoric references contained in certain messages and, although they can search the web for information, they do not always find the keys for interpreting the content of these messages (example 23). This type of disagreement mainly affects the sarcasm and stance level, leading to disagreements in the Toxicity category (non-toxic and toxic comments), as well as in the Toxicity_level, especially between levels 1 and 2 (i.e. between mildly and

Table 4 Summary of partial and total disagreements

File_name	Partial disagreement (%)	Total disagreement (%)
20170819_CR	9.35	2.51
20190512_MI	13.87	3.62
20190513_MI	13.50	3.67
20190519_SO	19.26	8.72
20190716_CR	14.78	5.31
20190919_MI	30.30	4.55
20200403_MI	10.52	1.81
20200424_MI	43.21	0.62
20200608_SO	17.55	3.80
20200618_MI	18.32	7.70
20200621_SO	12.87	2.33
20200622_MI	16.09	10.31
20200626_SO	14.13	4.00
20200705_SO	18.67	14.39
20200708_MI	22.29	5.49
20200715_CR	19.76	9.89
20200715b_CR	15.35	3.10
20200726_SO	17.76	1.60
20200831_CR	26.18	1.28
20200901_CR	17.20	2.33
Average	19.36	4.88

toxic comments). In the very toxic comments (Toxicity_level=3), toxicity is more explicitly conveyed and, therefore, easier to annotate.

23. (a) Los más valientes. Haz que pase!
(b) The bravest. Make it happen!

The first sentence in example (23) refers to the statement made by Manuela Carmena referring to the migrants who make it over the Melilla Wall as ‘the best and the bravest’. This information is not necessarily known by all annotators and the message can therefore be interpreted in different ways.

Another source of disagreement is the **ambiguity** of the messages, since the authors’ communicative intentions are often insufficiently clear and the comment can therefore have **multiple possible readings**. Ambiguity is mainly at the morpho-syntactic and semantic levels. For instance, in example (24) the disagreement is produced by the different way of understand *joder*, which can be interpreted as an interjection or a verb. As an interjection, the translation in English would be ‘And 40€fuck!’ (non-toxic content, the writer is surprised or angry), whereas if *joder* is a verb then the translation would be ‘And fuck for 40€(toxic content). In this example, the ambiguity is also related to the use of incorrect punctuation. The use of exclamation marks and a comma - *Y por 40€, !‘joder!* - would have helped in the interpretation of *joder* as an interjection. The incorrect use of punctuation and grammatical errors are easy to find in messages posted on social media.

24. Y por cuarenta euros joder, eso es menos que una oveja en la Edad Media! ‘And for forty fucking euros, that’s less than a sheep in the Middle Ages!

In contrast, ambiguity also occurs because the comment has a literal or figurative reading. In example (25), the comment can be interpreted literally, that is, it is a great success because the number of fifty-seven deaths is low compared to the 30,000 people who arrived in Spain (non-toxic content). However, the same comment can be interpreted sarcastically, if the intention of the writer is to emphasize that fifty-seven deaths are too few, because too many people have arrived alive (toxic content). This type of disagreement is mainly related to categories such as sarcasm and mockery. In these cases, ambiguity can only be resolved if the necessary context is available and that is not always the case.

25. Cincuenta y siete de los 33.000 que llegaron este año es un gran logro. ‘Fifty-seven out of the almost 33,000 who have arrived this year is a great success.’

In these cases, the ambiguity mainly affects the decision to label a comment toxic or non-toxic.

Finally, disagreements are caused by **human error**, such as tagging a comment as non-toxic and then assigning a Toxicity_level to the same comment. These disagreements are very easy to detect and solve.

Once the gold standard corpus was obtained, we calculated annotation consistency using Cronbach’s alpha (Cronbach, 1951) or coefficient alpha, which allows us to assess the internal reliability of the annotation of each feature (Table 5).[24] The average annotation consistency obtained was a Cronbach’s alpha index score of 0.772, which is an acceptable result with a value very close to good (Wadkar et al., 2016). The categories related to Stereotype and Target_group obtained the lowest consistency values, whereas Stance categories are the most clearly consistent categories. Stereotype and Target_group are closely related categories since a stereotype usually targets a specific group and, surprisingly, the annotators had some difficulties in deciding the target group, especially in cases in which reference was made to a political group, such as *fachas* (‘fascists’) or *podemitas*.[25] Stance categories were easier to identify, probably because some of the annotators had participated in the annotation with stance

[24] Cronbach’s alpha formula: $\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$. N=number of items; $\bar{c}$=average covariance between item pairs; $\bar{v}$=average variance. “The Coefficient alpha is used to assess the degree to which the items are internally consistent. A high coefficient indicates that the items are interrelated. More specifically, the performance on any one item can predict the performance on each of the remaining items” (Cronbach, 1951).

[25] Supporters of Podemos, a left-wing political party.

of another corpus (Taulé et al., 2017, 2018). The remaining categories, whose alpha values are very similar, are probably the most subjective categories and some of them are more likely to be expressed implicitly (for instance Sarcasm, Mockery, Intolerance and Toxicity categories) (Schmeisser-Nieto et al., 2022).

5.3 Hypotheses

The annotation of toxicity is based on the assumption that the category types, such as Stance, Target, Stereotype, Mockery, Insult, Improper_language, Aggressiveness and Intolerance, and their combination allows us to determine the degree of toxicity (Toxicity_level) in a more consistent way. Therefore, we are especially interested in discovering which combinations of categories are present at each Toxicity_level. This will allow us to identify the best features to define the toxicity of the comment and the intensity of this toxicity (see Sect. 6). In this subsection, we formulate the three hypotheses on which the selection of categories that we take into account for the annotation of toxicity is based.

H1 Primary comments (Comment_level=1), which refer to the news itself, tend to be less toxic, aggressive and intolerance, while secondary comments (Comment_level=2), which refer to a previous comment, tend to be more emotional, more personal and more toxic. Therefore, we are interested in observing the relationship between the Comment_level and Toxicity_level categories.

H2 Comments with a stance against the content of the news article, or of another comment, may present the contents in a more negative, more emotional and even more visceral way. Comments may be more sarcastic or may mock people, ideas or groups, and therefore may be more toxic than comments that share the same position or ideas presented in the news article or comment referred to. We are therefore interested in checking the relationship between the Negative_stance or Positive_stance and Toxicity_level categories, the Negative_stance or Positive_stance and Sarcasm categories and the Negative_stance or Positive_stance and Mockery categories.

H3 The presence of stereotypes (stereotyped prejudices) can lead to negative judgments or attitudes of intolerance that can in turn result in discrimination, marginalization and the exclusion of a specific group. In other words, stereotypes can make the contents of the comment more toxic. Therefore, we are interested in checking the relationship between the Stereotype and Toxicity_level categories, the Stereotype and Intolerance categories and the Stereotype and Target_group categories.

The annotation of the NewsCom-TOX corpus with all these categories will allow us to corroborate these hypotheses. In particular, the annotated corpus will allow us to analyze and evaluate which of the different proposed categories are most helpful for identifying toxicity and, especially, for determining the level of toxicity of the messages.

Table 5 Annotation consistency using Cronbach's alpha

Category	Cronbach's alpha (%) obtained when the feature is removed
Positive Stance	.783
Negative Stance	.781
Target person	.766
Target group	.700
Stereotype	.690
Sarcasm	.760
Mockery	.749
Insult	.760
Improper language	.767
Aggressiveness	.757
Intolerance	.757
Toxicity	.763
Toxicity level	.772
Average	.772

6 Results of the annotation of NewsCom-TOX

In this section, we present relevant quantitative data obtained from the annotation of the NewsCom-TOX corpus (Sect. 6.1). We also evaluate the hypotheses drawn up in Sect. 5.3 in relation to the data obtained from the annotation of the corpus (Sect. 6.2).

6.1 Quantitative analysis

The NewsCom-TOX corpus consists of 4359 comments distributed in 21 different files that correspond to 21 different news articles for which comments were posted. The number of comments for each file ranges from 65 to 359 per article. Table 6 shows the absolute and relative data for the comments annotated as toxic and their level of toxicity. The number of comments annotated with toxicity in the corpus is 1389, which represents 31.87% of the total of comments annotated. The percentage of comments annotated as toxic also varies ranging from 11.50% to 72.81% depending on the file.

Most of the comments annotated as toxic are mildly toxic (Toxicity_level=1) 71.99%, compared to 22.31% toxic (Toxicity_level=2) and 5.68% very toxic (Toxicity_level=3) comments (see Table 6). Moreover, in some files no comments were annotated as very toxic. However, it is worth noting that there are some files in which the difference between toxicity levels is smaller. This occurs in files in which the percentage of comments annotated as toxic is higher (>

Table 6 General corpus data

File _Name	Comments	Toxic comments	% Toxic comments	Toxic level 1	% Toxic level 1	Toxic level 2	% Toxic level 2	Toxic level 3	% Toxic level 3
20170819_CR	199	55	27.64	41	20.60	11	5.53	3	1.51
20190512_MI	359	119	33.15	73	20.33	34	9.47	12	3.34
20190513_MI	191	61	31.94	43	22.51	16	8.38	2	1.05
20190519_SO	172	54	31.40	42	24.42	10	5.81	2	1.16
20190919_MI	132	36	27.27	20	15.15	13	9.85	3	2.27
20190716_CR	320	233	72.81	171	53.44	51	15.94	11	3.44
20200403_MI	221	35	15.83	32	14.48	3	1.36	0	0.00
20200424_MI	162	51	31.48	43	26.54	6	3.70	2	1.23
20200608_SO	237	36	15.19	28	11.81	6	2.53	2	0.84
20200618_MI	65	27	41.54	10	15.38	13	20.00	4	6.15
20200621_SO	86	34	39.53	24	27.91	7	8.14	3	3.49
20200622_MI	194	65	33.51	44	22.68	18	9.28	3	1.55
20200626_SO	200	23	11.50	22	11.00	1	0.50	0	0.00
20200705_SO	132	61	46.21	38	28.79	18	13.64	5	3.79
20200708_MI	328	145	44.21	110	33.54	27	8.23	8	2.44
20200715_CR	273	94	34.43	65	23.81	24	8.79	5	1.83
20200715b_CR	129	31	24.03	23	17.83	6	4.65	2	1.55
20200725_CR	256	86	33.59	58	22.66	22	8.59	6	2.34
20200726_SO	125	24	19.20	20	16.00	4	3.20	0	0.00
20200831_CR	233	60	25.75	47	20.17	9	3.86	4	1.72
20200901_CR	345	59	17.10	46	13.33	11	3.19	2	0.58
Total	4359	1389 (31.87%)		1000 (71.99%)		310 22.31%)		79 (5.68%)	

Table 7 Distribution of categories by toxicity level in absolute and relative numbers

Toxicity level	Positive stance	Negative stance	Target person	Target group	Stereotype	Sarcasm	Mockery	Insult	Improper language	Aggressiveness	Intolerance
1	24	283	341	288	177	286	250	163	185	38	96
	2.40%	28.30%	34.10%	28.80%	17.70%	28.60%	25.00%	16.63%	18.50%	3.80%	9.60%
2	11	101	107	156	112	90	110	104	112	54	80
	3.55%	32.58%	34.52%	50.32%	36.13%	29.03%	35.48%	33.55%	36.%	17.42%	25.81%
3	1	32	30	50	28	10	34	54	45	30	33
	1.27%	40.51%	37.97%	63.29%	35.44%	12.66%	43.04%	68.35%	56.96%	37.97%	41.77%

40%). Therefore, we could conclude that the more toxic comments are posted to a news article, the higher the level of toxicity of comments.

Table 7 shows the distribution of categories by Toxicity_level, i.e. what type of category predominates at each Toxicity_level and, therefore, what categories best identify the level of toxicity. For instance, Insult, Target_group, Improper_language, Mockery and Intolerance categories, in this order, are by far those that best characterize Toxicity_level=3. These categories may also be considered the most explicit (except Mockery). The presence of insults implies the use of improper language, because these two categories are highly correlated. In fact, the Insult category appears at all levels of toxicity, but it is especially higher in very toxic comments. The difference with the other toxicity levels lies in the fact that, at the highest toxicity level, more insults or/and more offensive insults are used. Therefore, the assignment of the Toxicity_level depends, to a great extent, on the number and type of insults that a comment contains. A classification of insults considering their degree of offensiveness or hurtfulness would be helpful. Once again, we have to deal with subjectivity when interpreting their degree of offensiveness. The same thing happens, for instance, with categories such as Mockery and Intolerance, which are also identifiers of toxicity and are present in all the toxicity levels, and we will be classified at different toxicity levels depending on their level of Mockery or Intolerance. Although the Aggressiveness and Intolerance categories do not present such high percentages, they appear more frequently at Toxicity_level=3 than at level_2, and both are defining categories of Toxicity. In contrast, the two Target categories (person and group) appear more frequently at level_3, but with less difference with respect to level_2. It is noteworthy that the Target_person appears more frequently in mildly toxic comments than the Target_group.

It should also be noted that the Stereotype and Sarcasm categories are more frequently used in comments at Toxicity_level=2, although Stereotype has a very similar percentage at levels 2 and 3. Therefore, we could interpret this to mean that the appearance of a Stereotype or the use of Sarcasm, by themselves, do not increase Toxicity to the maximum level, but that we need other explicit elements for the degree of toxicity to increase. In the case of Sarcasm, we are also dealing with a very subjective category, which may also contain different degrees of criticism.

The low presence of the Positive_stance category is not surprising (to a certain extent) because comments are not usually made to support the news, but rather to criticise, question or oppose it. Therefore, the appearance of the Positive_stance is proportionally much lower than the Negative_stance.

Table 8 shows the combination of categories that appear most frequently at each toxicity level. For instance, Target_person and Sarcasm or Mockery is the most frequent combination in comments annotated as mildly toxic (level_1), followed by Negative_Stance and Target (independently of whether the comment is addressed to a person or a group). The combination of Target (person or group), Insult or Improper_language are the most frequent in those comments annotated as toxic (level_2), followed by Target_group, Stereotype and Intolerance. In contrast, the combination of Negative_Stance, Target (person or group), Insult or Improper_language and Intolerance categories is the most frequent in comments annotated as very toxic (level_3).

6.2 Evaluating the hypotheses

We drew up a contingency table to estimate the association between pairs of categories (Chi-squared test), that is, to observe whether these categories are independent or related, as well as their statistical significance and the corresponding odds ratio, in order to validate the hypotheses that we formulated at Section 5.3. The odds ratio (OR) “provides an estimate (with confidence interval) for the relationship between two binary (“yes or no”) variables” (Bland & Altman, 2000: p. 1468). We apply the OR to verify whether there exists a relation between our binary categories, for instance between Stereotype and Toxicity, thereby enabling us to confirm our initial hypotheses. Therefore, we will know how much more likely the occurrence of two specific categories is than their non-occurrence.

Table 9 shows the Chi-square value (χ^2), the statistical significance (p-value) and, whenever possible the corresponding OR, for each pair of categories. We will now review the hypotheses formulated taken into account the results displayed in Table 9.

Regarding hypothesis 1, in which we are interested in observing the relationship between the Comment_level and Toxicity_level categories, we observed that a direct relationship exists between them (χ^2 = 135.543 p-value < .05). However, the hypothesis cannot be validated because the level of toxicity is higher in primary comments than in secondary comments (i.e. comments referring to a previous comment). The percentage of secondary comments (64.95%) annotated as non-toxic is also higher than the percentage of primary comments (35.05%) (see Table 10). We should take a deeper look at the conversational threads of each comment, because we have only annotated whether the comments are primary or secondary, and we have annotated as secondary all the comments that are not primary regardless of the degree of nesting of the comment in the discourse thread. Therefore, this fact may explain the higher presence of secondary comments annotated as non-toxic. It would be interesting to analyze, in particular, those comments that generate the most debate because they are likely to contain more toxic comments. The relationship between Comment_level and Aggressiveness categories cannot be validated because this association is not statistically significant (χ^2 = .352 p-value > .05). However, there is an association between Comment_level and Intolerance categories (χ^2= 46.304 p-value < .05), but it is an indirect association (OR = .406), which means that the Intolerance category appears most frequently in secondary comments.

The relationship between Stance and Toxicity_level formulated in Hypothesis 2 can only be validated partially: (a) Negative_stance and Toxicity_level categories maintain a significant direct relationship (χ^2 = 136.640 p-value < .05), whereas we are unable to establish a statistically significant association between Positive_stance and Toxicity_level (χ^2 = 2.662 p-value > .05); (b) The relationship between Negative and Positive stance and Sarcasm categories cannot be established because the association is also not statistically significant (χ^2 = 1.943 p-value > .05 and χ^2 = 3.792 p-value > .05 respectively); (c) However, the relationship between Negative and Positive stance and Mockery categories is validated in both cases, though Negative_stance and Mockery categories show a direct positive association (OR = 2.538), whereas Positive_stance and Mockery

Table 8 The most frequent combinations of categories by Toxicity level.'/'stands for 'or'

Toxicity_level 1		
Target_person + Sarcasm/Mockery	230	23.00%
Negative_Stance + Target_(person/group)	185	18.50%
Negative_Stance + Improper language	71	7.10%
Toxicity_level 2		
Target_group + Insult/Improper_language	62	20.00%
Target_person + Insult/Improper_language	61	19.68%
Target_group + Stereotype + Intolerance	46	14.84%
Toxicity_level 3		
Negative_Stance + Target_(person/group)+Insult/Improper_language+Intolerance	11	13.92%
Target_group + Insult + Improper_language + Mockery	9	11.39%
Target_group + Stereotype + Mockery + Intolerance	7	8.86%

For instance, 'Target_Person + Sarcasm/Mockery' means that the combinations Target_Person and Sarcasm or Target_Person and Mockery appears 230 times in comments

show a negative association (OR = 0.214), which means that if the Stance is annotated as positive, Mockery tends not to be found.

Finally, Hypothesis 3, which claims that (a) stereotypes (stereotyped prejudices) can make the contents of the comments more toxic, and (b) stereotypes can lead to attitudes of intolerance, is also validated (χ^2 = 666.149 p-value < .05 and χ^2= 563.984 p-value < .05 OR = 15.468, respectively) (Table 10). These relationships present high association values. We observe a strong level of association between both categories: when Stereotype does not appear, Toxicity and Intolerance do not appear in the majority of cases. However, the percentage of stereotypes increases as the level of toxicity rises and we also observe that the presence of stereotypes also increases in line with the percentage of examples annotated with Intolerance (Table 7). As expected, the claim that stereotypes are attributed to a group is also validated, as shown by the high level of association between these features (χ^2 = 1532.392 p-value <.05 OR= 49.856), which, indeed, present the highest value of the categories analyzed. The data also show that the relationship between Target (regardless of whether it is a specific person or a group) and Toxicity categories is also statistically significant. The level of toxicity tends to rise when the comments are addressed to a target, and especially when the target is a group (Target_group) (Table 7).

We can also observe that the association of the remaining categories -Target, Mockery, Insult, Improper_language, Aggressiveness and Intolerance- with the Toxicity_level is statistically significant (χ^2 ranging from 587.128 to 1034.333 and p-values lower than.05), having the highest associations with Insult, Target and Mockery.

Table 9 Association between pairs of categories

Category pair	χ^2	p	OR
Comment_level/Toxicity_level	135.543	<.05	(a)
Comment_level/Aggressiveness	.352	.553 (b)	(b)
Comment_level/Intolerance	46.304	<.05	.406 (b)
Positive_Stance/Toxicity_level	2.662	.447 (b)	(a) (b)
Negative_Stance/Toxicity_level	136.640	<.05	(a)
Negative_stance/Sarcasm	1.943	.163 (b)	(b)
Positive_stance/Sarcasm	3.792	.052 (b)	(b)
Negative_stance/Mockery	76.742	<.05	2.538
Positive_stance/Mockery	8.400	<.05	.214 (c)
Stereotype/Toxicity_level	666.149	<.05	(a)
Stereotype/Intolerance	563.984	<.05	15.468
Stereotype/Target_person	.251	.616 (b)	(b)
Stereotype/Target_group	1532.392	<.05	49.856
Target_person/Toxicity_level	717.441	<.05	(a)
Target_group/Toxicity_level	1025.969	<.05	(a)
Stereotype/Toxicity_level	666.149	<.05	(a)
Sarcasm/Toxicity_level	662.038	<.05	(a)
Mockery/Toxicity_level	856.701	<.05	(a)
Insult/Toxicity_level	1034.333	<.05	(a)
Improper_language/Toxicity_level	670.927	<.05	(a)
Aggressiveness/Toxicity_level	642.163	<.05	(a)
Intolerance/Toxicity_level	587.128	<.05	(a)

(a) Indicates that the OR cannot be applied because one of the categories has more than two values (b) Indicates that the relation between categories is not statistically significant (p > .05). The relation between categories is statistically significant when p < .05. (c) Indicates that OR<1, which means that the association is negative, whereas OR>1 indicates that the association is positive (i.e. the presence of the first category is associated with a higher occurrence of the second category)

Table 10 Percentage of Comment_level_1 and Comment_level_2 by Toxicity_leve_l

	Comment_level_1	Comment_level_2	Total
Toxicity_level_0	1041	1929	2970
	35.05%	64.95%	
Toxicity_level_1	529	471	1000
	52.90%	47.10%	
Toxicity_level_2	161	149	310
	51.94%	48.06%	
Toxicity_level_3	51	28	79
	64.56%	35.44%	
Total Comments	1782	2,577	4359
	40.88%	59.12%	

7 Conclusions

In this article we have described the methodology applied to the annotation of the NewsCom-TOX corpus, which is the first corpus annotated with different levels of toxicity in comments written in the Spanish language posted to online news articles. The NewsCom-TOX corpus consists of 4359 comments, out of which, 1389 are toxic (31.87%) and 2970 (68.13%) are not toxic. Each comment has been annotated with 9 different linguistic categories (which correspond to 11 different labels) taking into account contextual information (the discourse thread) and the extra-linguistic context (news articles and images), which have been used to determine whether or not the comment was toxic and its degree of toxicity. The majority of comments annotated as toxic are mildly toxic (71.99%), compared to 22.31% toxic and 5.68% very toxic comments. The inter-annotator agreement tests conducted to test the reliability of the annotation obtained a total average of observed agreement of 81.84% (0.59 α) for Toxicity and 73.58% (0.54 α) for the Toxicity_level, which ensures annotation reliability. Although these results are quite acceptable, they also evidence the complexity of the task carried out: we can annotate and apply the criteria defined consistently, but there is always a part of the interpretation of comments that is inevitably subjective. This difficulty is even more evident when some of the categories used to determine the level of toxicity also require a subjective interpretation, such as Sarcasm, Intolerance and Aggressiveness, and many of them also involve a certain gradation, such as, Insult, Improper language, Mockery, Intolerance and Sarcasm, which may be presented in a more or less offensive degree. Furthermore, it is very difficult to avoid the bias of annotators, even though the annotation has been performed by the same annotators, all of whom participated in the definition (or improvement) of the guidelines and in the discussion sessions held once a week to solve disagreements. Classifying contents into discrete categories based on a scale -mildly toxic, toxic and very toxic- is very difficult.

The results allow us to conclude that the number of categories and especially the type of categories contained in the comments are crucial to classify the toxicity level of comments. It is worth noting that categories such as Insult, Improper_language and Mockery can convey Toxicity more explicitly than Stereotypes, Sarcasm, Intolerance and Aggressiveness, in which it is conveyed more implicitly. The presence of stereotypes, which are almost always targeted to a group, increases the level of toxicity. It would be interesting to analyse whether different types of stereotypes (for instance racial stereotypes related to crimes) entail more toxicity than others (for instance racial stereotypes related to benefits). In fact, comments in which the Target person or group is explicit are mostly toxic, especially when the Target is a group, and their level of toxicity rises when Insults and Mockery, Intolerance and Aggressiveness appear. It would also be interesting to explore in a future work the possibility to ponderate each category, giving more weight to those categories that are more prone to a higher level of toxicity, with the aim of improving or fine-tuning the level of toxicity. We can also conclude that, in order to assign each of these categories and be able to interpret the global meaning that comments convey, it is definitely essential to take the context into account, that is, the discourse thread and the extra-linguistic context. In future work we would like to go deeper and take into account the degree of nesting of comments in the discursive thread. Finally, we are enlarging the NewsCom-TOX corpus with more comments extracted from different news articles in order to be used as training and test corpus for developing models to automatically detect toxicity and the level of toxicity. This corpus can be very useful for the linguistic analysis of toxicity but also for studying each specific linguistic category annotated more deeply and the linguistic patterns in which they occur. This corpus has been used in the DETOXIS shared task (Taulé et al., 2021), which was hold in the IberLEF 2021 workshop.

Appendix

See Table 11.

Table 11 Distribution of comments per topic in the NewsCom-TOX corpus CR=Crime; MI=Migration; SO=Society - Original titles in Spanish

File_name (Date_Topic)	Comments	News article title (Source link)
20171908_CR	199	Los investigadores sitúan al nuevo imán de Ripoll al frente del grupo terrorista (https://www.eldiario.es/) + (https://www.meneame.ne/)
20190512_MI	359	Al menos 63 muertos al naufragar en Mauritania una patera que iba a Canarias (https://www.elpais.com) + (https://www.forocoches.com)
20190513_MI	191	El Gobierno pide a la Iglesia que acoja a los refugiados que rechaza la izquierda (https://www.abc.es)
20190519_SO	172	Por qué las víctimas de violencia machista pueden encadenar/prompt relaciones tóxicas (https://www.elpais.com)
20190716_CR	320	Un ’cortacabezas’ en la patera (https://www.elmundo.es) + (https://www.forocoches.com)
20190919_MI	132	El Gobierno aragonés busca siete chalés para alojar menas por 300.000 euros al mes (https://www.abc.es) + (https://www.forocoches.com)
20200403_MI	221	Se necesitan 150.000 personas para trabajar en el campo (https://www.elmundo.es) + (https://www.forocoches.com)
20200424_MI	162	Más de 5.000 euros por escapar de España en patera (https://www.elmundo.es)
20200608_SO	237	David Cantero se hace viral por 27 segundos: sus palabras sobre los inmigrantes en España (https://www.ecoteuve.eleconomista.es)
20200618_SO	65	Los jóvenes ocupantes de Premiá acumulan más de veinte antecedentes por robos violentos (https://www.lavanguardia.com) + (https://www.meneame.net)
20200621_SO	86	Subdelegación del Gobierno ordena investigar a Javier Negre por el vídeo con un vendedor ambulante en Galicia (https://www.eldiario.es)
20200622_MI	194	Un joven magrebí recibe una paliza tras un robo con violencia a una anciana (https://www.niusdiario.es) + (https://www.meneame.net)
20200626_SO	200	Lacasa, obligada a aclarar que los Conguitos son cacahuetes y no personas (https://www.libremercado.com) + (https://www.meneame.net)
20200705_SO	132	El cliente de un bar a la policía: “Con estas putas negras no quiero compartir terraza” (https://www.ultimahora.es)
20200708_MI	328	Unidas Podemos propone regularizar a los 600.000 inmigrantes sin papeles que estaban en España cuando se decretó el estado de alarma (https://www.elmundo.eswww.elmundo.es)

Table 11 (continued)

File_name (Date_Topic)	Comments	News article title (Source link)
20200715_CR	273	Le amputan un dedo a machetazos al quejarse del ruido que hacen en un parque de Madrid (https://elcaso.elnacional.cat)
20200715b_CR	129	The Le amputan un dedo a machetazos al quejarse del ruido que hacen en un parque de Madrid (https://www.extremadura7dias.com) + (https://www.meneame.net)
20200725_CR	256	Dos menores estrangulan a una mujer para robarle el móvil en un túnel de Batán: "Me dejaron inconsciente" (https://www.abc.es)
20200726_SO	125	La Fiscalía de Málaga pide el cierre del medio Alerta digital por incitación al odio (https://www.eldiario.es) + (https://www.meneame.net)
202008031_CR	233	Al menos cuatro detenidos por abusar de una menor y dar palizas en el parque Wifi (https://www.ultimahora.es) + (https://www.meneame.net)
20200901_CR	345	Trump defiende al joven de 17 años detenido por matar a dos manifestantes en Wisconsin: "Se cayó y le atacaron" (https://www.20minutos.es) + (https://www.mename.net)
Total	4359	21 articles

Acknowledgements The work has been carried out in the framework of the following projects: FairTransNLP-Language: Analysing Toxicity and Stereotypes in Language for Unbiased, Fair and Transparent Systems (PID2021-124361OB-C33) funded by Ministerio de Ciencia e Innovación (Spain) and co-funded by the European Regional Development Fund (FEDER), CLiC SGR (2027SGR341), funded by AGAUR (Generalitat de Catalunya) and STERHEOTYPES project (Challenges for Europe), funded by Fondazione Compangia di San Paolo.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Abercrombie, G., Basile, V., Tonelli, S., Rieser, V., & Uma, A. (2022). (eds) Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, *European Language Resources Association, Marseille, France,*https://aclanthology.org/2022.nlperspectives-1.0.

Akhtar, S., Basile, V., & Patti, V. (2021). Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint*arXiv:2106.15896.

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

Álvarez-Carmona, MÁ., Guzmán-Falcón, E., Montes-Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: *Notebook papers of 3rd SEPLN workshop on Evaluation of human language technologies for Iberian languages (IberEVAL), Seville, Spain*, vol 6.

Aragón, M. E., Jarquín-Vásquez, H. J., Montes-Gómez, M., Escalante, H.J., Pineda, L.V., Gómez-Adorno, H., Posadas-Durán, J.P., & Bel-Enguix, G. (2020). Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In: *IberLEF@ SEPLN,* pp 222–235.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596.

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., Rosso, P., Sanguinetti, M. et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: *13th International Workshop on Semantic Evaluation, Association for Computational Linguistics,* pp 54–63.

Beukeboom, C. J., & Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the Social Categories and Stereotypes Communication (SCSC) framework. *Review of Communication Research*. https://doi.org/10.12840/issn.2255-4165.017

Bland, J., & Altman, D. (2000). Statistics notes. The odds ratio. *British Medical Journal, 320*(7247), 1468.

Bosco, C., Lai, M., Patti, V., Pardo, F. M. R., & Paolo, R. et al. (2016). Tweeting in the debate about catalan elections. In: *Workshop on Emotion and Sentiment Analysis, European Language Resources Association (ELRA)*, pp 67–70.

Caselli, T., Basile, V., Mitrovic, J., Kartoziya, I., Granitzer, M. (2020). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, *European Language Resources Association,* pp 6193–6202, https://www.aclweb.org/anthology/2020.lrec-1.760/

Cignarella, A. T., Lai, M., Bosco, C., Patti, V., & Paolo, R. et al. (2020). Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets. In: *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Ceur*, pp 1–10.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In: *Proceedings of the International AAAI Conference on Web and Social Media,* vol 11

Díaz-Torres, M. J., Morán-Méndez, P. A., Villasenor-Pineda, L., Montes-y Gómez, M., Aguilera, J., & Meneses-Lerín, L. (2020). Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France,* pp 132–136, https://aclanthology.org/2020.trac-1.21.

Dynel, M. (2021). Desperately seeking intentions: Genuine and jocular insults on social media. *Journal of Pragmatics, 179*, 26–36. https://doi.org/10.1016/j.pragma.2021.04.017

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In: *12th International AAAI Conference on Web and Social Media, ICWSM 2018, AAAI Press,* pp 42–51.

Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics, 30*(1), 95–101.

Farias, D. H., & Rosso, P. (2017). Chapter 7 - irony, sarcasm, and sentiment analysis. In F. A. Pozzi, E. Fersini, E. Messina, & B. Liu (Eds.), *Sentiment Analysis in Social Networks* (pp. 113–128). Boston: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-804412-4.00007-3

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. In: *CEUR Workshop Proceedings, vol 2150*, pp 214–228, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054936434 &partnerID=40 &md5=6e6b1965d972a7e6c220166577253324, cited By 33.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*. https://doi.org/10.1145/3232676

Fortuna, P., Soler, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In: *Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France,* pp 6786–6794, https://aclanthology.org/2020.lrec-1.838.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In: *AAAI International Conference on Web and Social Media (ICWSM).*

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2020). The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics, 4*(2), 155–190.

Krippendorff. K. (2004). Content analysis: An introduction to its methodology (2 nd) Thousand oaks.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018),* pp 1–11.

Lai, M., Patti, V., Ruffo, G., & Rosso, P. (2018). Stance evolution and twitter interactions in an italian political debate. In: *International Conference on Applications of Natural Language to Information Systems, Springer,* pp 15–27.

Mall, R., Nagpal, M., Salminen, J., Almerekhi, H., Jung, S. G., & Jansen, B. J. (2020). Four types of toxic people: Characterizing online users' toxicity over time. In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, Association for Computing Machinery, New York, NY, USA, NordiCHI '20*https://doi.org/10.1145/3419249.3420142

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In: *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp 31–41.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In: *Proceedings of the 25th international conference on world wide web*, pp 145–153.

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American Constitution, 3*(2), 1277–1279.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). Toxicity detection: Does context really matter? In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* pp 4296–4305, https://doi.org/10.18653/v1/2020.acl-main.396.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors, 19*(21), 4654.

Plaza-delArco, F. M., Casavantes, M., Escalante, H. J., Martín-Valdivia, M. T., Montejo-Ráez, A., Montes-Gómez, M., Jarquín-Vásquez, H., & Villaseñor-Pineda, L. (2021). Overview of meoffendes at iberlef 2021: Offensive language detection in Spanish variants. *Procesamiento del Lenguaje Natural, 67*, 183–194.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* pp 1–47

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., & Donoso, T. (2021). Overview of exist 2021: Sexism identification in social networks. *Procesamiento del Lenguaje Natural, 67*, 195–207.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Gonzalo, J., Spina, D., & Rosso, P. (2022). Overview of exist 2022: Sexism identification in social networks. *Procesamiento del Lenguaje Natural, 69*, 229–240.

Salminen, J., (1, Chowdhury, S., Jung, S. G., Jansen, B., Hopf, M., Almerekhi, H. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10(1), http://sire.ub.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true &db=edselc &AN=edselc.2-52.0-85077201223 &lang=es &site=eds-live.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018a). An italian twitter corpus of hate speech against immigrants. In: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*

Sanguinetti, M., Poletto, F, Bosco, C., Patti, V., & Stranisci, M. (2018b) .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,*https://aclanthology.org/L18-1443

Schmeisser-Nieto, W., Nofre, M., & Taulé, M. (2022). Criteria for the annotation of implicit stereotypes. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* pp 753–762.

Schmidt, A., & Wiegand, M. (2017a). A survey on hate speech detection using natural language processing. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain*, pp 1–10, https://doi.org/10.18653/v1/W17-1101, https://aclanthology.org/W17-1101.

Schmidt, A., & Wiegand, M. (2017b). 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.

Sánchez-Junquera, J., Chulvi, B., Rosso, P., & Ponzetto, S. P. (2021). How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*. https://doi.org/10.3390/app11083610

Taulé, M., Martí, M. A., Rangel, F. M., Rosso, P., Bosco, C., & Patti, V. et al. (2017). Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In: *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017, CEUR-WS,* vol 1881, pp 157–177.

Taulé, M., Pardo, F. M. R., Martí, M. A., & Rosso, P. (2018). Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. In: *IberEval@ SEPLN,* pp 149–166

Taulé, M., Ariza, A., Nofre, M., Amigó, E., & Rosso, P. (2021). Overview of the detoxis task at iberlef-2021: Detection of toxicity in comments in spanish. *Procesamiento del Lenguaje Natural* 67

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., & Poesio, M. (2021). Semeval-2021 task 12: Learning with disagreements. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics,* pp 338–347.

Vidgen, B., & Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE, 15*(12), 1–32. https://doi.org/10.1371/journal.pone.0243300

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In: *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018),*http://sire.ub.edu/login?url=http://

search.ebscohost.com/login.aspx?direct=true &db=edsair &AN=edsair.narcis........d3a04b3fc898d65718f3e30330f11b71 &lang=es &site=eds-live.

Wadkar, S. K., Singh, K., Chakravarty, R., & Argade, S. (2016). Assessing the reliability of attitude scale by Cronbach's alpha. *Journal of Global Communication, 9*, 113–117.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: *Proceedings of the NAACL student research workshop,* pp 88–93

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In: *Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguisticss,* pp 78–64, https://aclanthology.org/W17-3012.pdf

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In: *WWW '17: Proceedings of the 26th International Conference on World Wide Web*, pp 1391–1399, https://doi.org/10.1145/3038912.3052591

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics,* pp 1425–1447, https://aclanthology.org/2020.semeval-1.188.