IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 January 2023, accepted 4 March 2023, date of publication 20 March 2023, date of current version 30 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3258973

RESEARCH ARTICLE

DISCLAIMER: This article contains examples of offensive and abusive language as a necessary part of illustrating its research findings.

Assessing the Impact of Contextual Information in Hate Speech Detection

JUAN MANUEL PÉREZ[1], FRANCO M. LUQUE[2,3], DEMIAN ZAYAT[4], MARTÍN KONDRATZKY[5], AGUSTÍN MORO[3,6], PABLO SANTIAGO SERRATI[3,7], JOAQUÍN ZAJAC[3,8], PAULA MIGUEL[3,7], NATALIA DEBANDI[9], AGUSTÍN GRAVANO[3,10,11], AND VIVIANA COTIK[1,12]

[1]Instituto de Ciencias de la Computación, CONICET, Universidad de Buenos Aires (UBA), Buenos Aires 1053, Argentina
[2]Facultad de Astronomía, Matemática y Física, Universidad Nacional de Córdoba, Córdoba 5000, Argentina
[3]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires 1033, Argentina
[4]Facultad de Derecho, Universidad de Buenos Aires, Buenos Aires 1053, Argentina
[5]Facultad de Filosofía y Letras, Universidad de Buenos Aires, Buenos Aires 1053, Argentina
[6]Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil 7300, Argentina
[7]Instituto de Investigaciones Gino Germani, Facultad de Ciencias Sociales, Universidad de Buenos Aires, Buenos Aires 1053, Argentina
[8]Escuela Interdisciplinaria de Altos Estudios Sociales, Universidad de San Martín, San Martín 1650, Argentina
[9]Universidad Nacional de Río Negro, Rio Negro 8500, Argentina
[10]Escuela de Negocios, Universidad Torcuato Di Tella, Buenos Aires 7350, Argentina
[11]Laboratorio de Inteligencia Artificial, Universidad Torcuato Di Tella, Buenos Aires 7350, Argentina
[12]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires 1053, Argentina

Corresponding author: Juan Manuel Pérez (jmperez@dc.uba.ar)

This work was supported by the interdisciplinary projects evaluated and funded by Universidad de Buenos Aires under Grant PIUBAMAS-2020-3 and Grant PIUBA-2022-04-02.

ABSTRACT Social networks and other digital media deal with huge amounts of user-generated contents where hate speech has become a problematic more and more relevant. A great effort has been made to develop automatic tools for its analysis and moderation, at least in its most threatening forms, such as in violent acts against people and groups protected by law. One limitation of current approaches to automatic hate speech detection is the lack of context. The spotlight on isolated messages, without considering any type of conversational context or even the topic being discussed, severely restricts the available information to determine whether a post on a social network should be tagged as hateful or not. In this work, we assess the impact of adding contextual information to the hate speech detection task. We specifically study a subdomain of Twitter data consisting of replies to digital newspapers posts, which provides a natural environment for contextualized hate speech detection. We built a new corpus in Spanish (*Rioplatense variant*) focused on hate speech associated to the COVID-19 pandemic, annotated using guidelines carefully designed by our interdisciplinary team. Our classification experiments using state-of-the-art transformer-based machine learning techniques show evidence that adding contextual information improves the performance of hate speech detection for two proposed tasks: binary and multi-label prediction, increasing their Macro F1 by 4.2 and 5.5 points, respectively. These results highlight the importance of using contextual information in hate speech detection. Our code, models, and corpus has been made available for further research.

INDEX TERMS NLP, text classification, hate speech detection, contextual information, Spanish corpus, COVID-19 hate speech.

I. INTRODUCTION

Hate speech can be described as speech containing denigration and violence towards an individual or a group of individuals, based on certain characteristics protected by international treaties, such as gender, race, language, and others [1]. In recent years, this type of discursivity became problematically relevant due to its intensity and its prevalence on social media. The exposure to this phenomenon has been associated with stress and depression of victims [2], and also

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

to the settle of a hostile and dehumanizing environment for immigrants, sexual and religious minorities, as well as other vulnerable groups [3]. Adding to the psychological effects, one of the most worrying aspects of hate speech on social media is its relationship with violent acts against members of these groups, such as the "Unite the Right" attacks at Charlottesville [4], the Pittsburgh synagogue shooting [5], and the Rohingya genocide at Myanmar [6], [7], among others. As a result, states and supranational organizations such as the European Union have enacted legislation that urges social media companies to moderate and eliminate discriminatory content, with a particular focus on that encouraging physical violence [8].

During the COVID-19 pandemic (2020-2021), a dramatic increase in the prevalence of hate speech has been seen, featuring targets such as Chinese, Asian, and Jews, among other nationalities and minorities blamed for the spread of the virus or the increase in social inequalities [9]. The dissemination of fake news related to conspiracy theories and other types of misinformation [10], [11] has been linked to an increase in violence against members of those groups [9].

In recent years, the need of the analysis and moderation of hate speech, at least in its most threatening forms, has increased and a great effort in research and development of automatic tools to address it has been made [12], [13], [14], [15]. From Natural Language Processing (NLP) perspective, hate speech detection can be thought of as a text classification task: given a text document generated by a user (i.e., a post in a social network), would be possible to predict whether or not it contains hateful content [14]. Additionally, other features, such as whether the text contains a call to take some violent action or not, if the message is directed against an individual or a group, or which characteristics are at the cause of the attack [16] among other possibilities, could be explored and analysed.

One limitation in current approaches on automatic hate speech detection is the lack of context. Many studies and resources work with data without any kind of context - i.e., isolated user messages with no information about the conversational thread or even the topic being discussed- [17]. This situation creates a limitation on the available information to detect if a comment is hateful or not , given that an expression can be injurious in certain contexts, but not in others.

Another limitation for hate speech detection is that most of resources are built in English, restricting the research and its applicability in other languages [14], [15]. While there are some datasets for hate speech detection in Spanish [16], [18], [19], to the best of our knowledge, none is related to the COVID-19 pandemic, which shows distinctive features and targets in comparison to other hate speech events.

In this paper, we address the issues described above regarding hate speech detection: 1) we consider **finer-grained** distinctions that go beyond a binary detection of hateful vs. non-hateful speech, such as the identification of attacked characteristics and the detection of calls to action; 2) we study the impact of adding **contextual information** to classification problems, and 3) we address the problem in **Spanish**, a language with relatively few resources available for this task. We are especially interested in the second issue, being the usefulness of contextual information the main research question in this work.

For these purposes, we built a dataset based on user responses to posts from Argentinian digital newspapers on Twitter. This subdomain of content in social networks (i.e., responses to news posts) is particularly interesting because it provides a natural context for the discussion (meaning the debate on news) while also replicating the interactions of a news forum. We collected data from news in Spanish related to the COVID-19 pandemic and a sample of the dataset was annotated by Spanish native speakers. As a plus, our dataset comes from the *Rioplatense* Spanish dialect,[1] which adds to neutral Spanish its own particularities and expressions of hate speech in a distinctive way. Classification experiments using state-of-the-art techniques based on *BETO* [20], a Spanish version of BERT (Bidirectional Encoder Representations from Transformers) [21], show evidence that adding context improves the performance in hate speech detection both in a binary setting (predicting the presence or absence of hate speech) and in a fine-grained setting (predicting attacked characteristics and whether there is a call to action or not). These results highlight the importance of contextual information for hate speech detection. Figure 1 provides a graphical, high-level overview of the work discussed in this paper.

Our contributions are the following:

1) We describe the collection, curation, and annotation process of an original corpus for hate speech detection based on user responses to news posts from media outlets on Twitter. This dataset is in Spanish (in its *Rioplatense* variety[2]) and focuses on hate speech associated with COVID-19 pandemic.
2) Through a series of classification experiments using state-of-the-art techniques, we show evidence that including contextual information improves the performance of hate speech detection, both in binary and fine-grained settings.
3) We make our code, models, and the annotated corpus available[3] for further research.

The paper is organized as follows: In Section II previous work for automatic hate speech detection is reviewed. Section III discusses the definition of hate speech used in this work, along with the targeted groups and the features of interest for this kind of hate speech. Section IV describes the process performed to collect data, curate and annotate a sample and build our corpus, which is later used in Section V to conduct our classification experiments. Section VII discusses

[1] To the best of our knowledge, no dataset exists for this variant.

[2] *Rioplatense* Spanish —or *Rioplatense* Castilian— is a variety of Spanish spoken mainly in and around the Río de la Plata Basin of Argentina and Uruguay.

[3] Code can be found at https://github.com/finiteautomata/contextualized-hatespeech-classification, and models and datasets at https://huggingface.co/piuba-bigdata

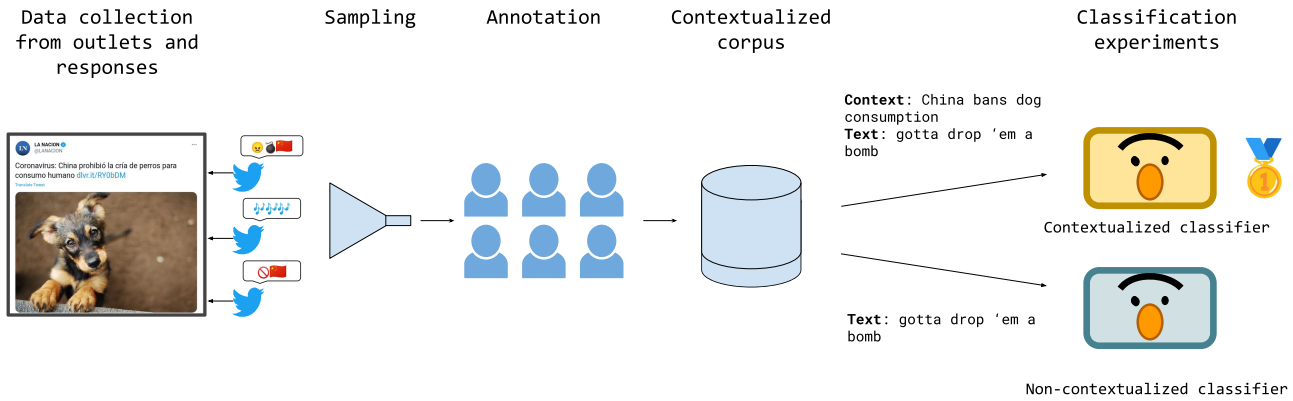


FIGURE 1. Work overview. The process starts with the collection of data from Twitter. A sampling procedure is designed and applied to achieve a balanced proportion of attacked characteristics. Then, the dataset is annotated by native speakers following carefully designed annotation guidelines. The annotated corpus is used to train and evaluate models for hate speech detection, both as a binary and a multi-label classification task. Our experiments reveal that contextualized models outperform non-contextualized ones.

the results and Section VIII draws conclusions and outlines possible future work.

II. PREVIOUS WORK

Hate speech has attired attention in recent years, with literature from legal and social sciences domains studying its definition and classification [22], the elements that enable its identification, and its rapport with debates on freedom of expression and human rights [1], [23]. The automatic detection of this kind of speech is usually addressed as a classification task, and it is related to a family of other tasks such as detecting cyberbullying, offensive language, abusive language, toxic language, among others. Reference [24] proposes a typology of these related tasks by asking if the offensive content is directed to a specific entity or group, and whether the content is explicit or implicit.

There is a plethora of resources for automatic detection of hate speech. Interested readers can refer to [17] for an extensive review of datasets addressing this task. Nevertheless, Spanish corpora are scarce, despite being Spanish the worldwide second language in number of native speakers and one of the most used languages in social media [25]. To the best of our knowledge, all available datasets in Spanish have been published in the context of shared tasks. Reference [19] presented a ∼4k Twitter dataset for the Automatic Misogyny Identification (AMI) shared task (IberEval 2018[4]). The MEX-A3T task (IberEval 2018 and IberLEF 2019[5]) included a dataset of ∼11k Mexican Spanish tweets annotated for aggressiveness [26], [27]. Reference [16] published a ∼6.6k tweets dataset annotated for misogyny and xenophobia, in the context of the HatEval challenge (SemEval 2019[6]).

In the context of the COVID-19 pandemic, a spike in the incidence of hate speech in social networks has been registered [28]. Some works have addressed its distinctive features, studying hateful dynamics in social networks [29] and also generating specific resources for the analysis and identification of this kind of problematic behavior [30]. Reference [31] describes a work-in-progress of this research on hate speech analyzing tweets in Spanish linked to newspaper articles on the COVID-19 pandemic.

Regarding the techniques applied for our specific task, classic machine learning techniques such as handcrafted features and bags of words over linear classifiers have been applied [12], [32], [33]. Lately, however, deep learning techniques such as recurrent neural networks or —more recently— pre-trained language models have become state-of-the-art [34], [35], [36], [37], [38], [39]. In spite of the great results achieved by those methods, [40] arises a question mark on some of them, suggesting that they may be subject of possible cases of overfitting. Reference [41] analyzes the currently available Spanish pre-trained models for hate speech detection tasks.

Since the appearance of GPT (Generative Pre-trained Transformer) [42] and BERT [21], pre-trained language models based on transformers [43] have become state-of-the-art for most NLP tasks. These techniques use a transfer-learning approach, by first pre-training a large language model (thus their name) on a big corpus, and then fine-tuning it for a specific task (e.g. sentiment analysis, question answering, or hate speech detection) [42], [44]. This approach has replaced previous deep learning architectures once developed for most NLP tasks, which used to be based on recurrent neural networks and word embeddings [45], [46].

Pre-trained models have been built for different languages, and also for different domains (such as biomedical [47] or legal domain [48]) and text sources (such as Twitter [49] and other social networks). In particular, Spanish pre-trained models include BETO [20], BERTin [50], RoBERTA-es [51] and RoBERTuito [52]. Reference [53] review BERT-based language models for different tasks and languages.[7]

[4]IberEval 2018: https://sites.google.com/view/ibereval-2018?pli=1
[5]IberLEF 2019: https://sites.google.com/view/iberlef-2019/
[6]SemEval 2019: https://alt.qcri.org/semeval2019/

[7]Note that the names BETO, BERTin, RoBERTA, and RoBERTuito are not acronyms, but alterations of the original name BERT.

Few prior studies have incorporated some context to user comments in hate speech detection. Reference [54] analyze the impact of adding context to the task of hate speech detection in a dataset of comments from the Fox News website. As mentioned by [55], this study has room for improvement: the dataset is rather small, with around 1.6k comments extracted from 10 news articles only; its annotation process was mainly performed by just one person; and some of its methodologies, such as including the name of the user as a predictive feature, are subject of discussion. Reference [56] built a dataset of comments taken from the Al Jazeera website[8] and annotated them along with the title of the article, but without including the entire thread of replies.

Reference [55] analyze the impact of adding context to toxicity detection task. They find that, while humans seem to leverage conversational context to detect toxicity, the trained classification models were not able to improve their performance considerably by adding context. Following up, [57] labeled each message with its ''context sensitiveness'', measured as the difference between two groups of annotators: those who have seen the context, and those who have not. With this, they observed that classifiers improve their performance on comments which are more sensitive to context.

Further, [58] explores some opportunities to incorporate richer information sources into the toxicity detection task, such as the interaction history between users, some social context, and other external knowledge bases. Reference [59] poses some questions and challenges regarding the detection of implicit toxicity — that is, some subtle forms of abusive language not expressed as strong language or insults.

Summing up, BERT-based models are state-of-the-art for this type of classification tasks; there have been various attempts to include context in different ways and with dissimilar success; there are relatively few studies on Spanish data; and hate speech detection has typically been addressed as a binary task, making no distinction among the attacked characteristics or calls-to-action. In the present work, we assess the usefulness of adding context, we work with BERT-based models, on Spanish data, and address both binary and fine-grained classification tasks.

III. DEFINITION OF HATE SPEECH

We say a comment involves hate speech if it contains statements of an intense and irrational disapproval and hatred against an individual or a group of people because of its identification with a group protected by domestic or international laws [1]. Protected traits or characteristics include color, race, national or social origin, language, gender identity, and sexual orientation, among others.

Hate speech could manifest explicitly as direct insults, slurs, celebrations of crimes, incitements to take action against an individual or group, or implicitly in more subtle ways and veiled expressions such as in ironic content. Following this definition, we consider that an insult or aggression is not enough to constitute hate speech; it is necessary to make an explicit or implicit appeal to at least one feature protected by law.

[8]https://www.aljazeera.com/

TABLE 1. Characteristics considered in this work. Short names are used throughout the paper to refer to these groups.

Short name	Hate speech against ...
WOMEN	women
LGBTI	gay, lesbian, bisexual, transgender, intersexual
RACISM	race, skin color, language, or national identity
CLASS	socioeconomic status
POLITICS	political affiliation or ideology
APPEARANCE	fat people, old people, or other aspect-based features
CRIMINAL	criminals and persons in conflict with law
DISABLED	phisical and mental disabilities or health affections

For international law, hate speech has an extra element that differentiates it from other offensive behavior: the promotion of violent actions against its targets. However, the NLP community does not usually require this ''call-to-action'' when identifying hate speech. In the present work, we adopt this latter view, and we explicitly express when we refer to calls to action.

Several characteristics are taken into account in this work. For their selection, we take into account the definition of discrimination from international human rights treaties, which refers to discrimination motivated by race, color, sex, language, religion, political, or other opinions, national or social origin, property, birth or other statuses [60]. So, in addition to misogyny and racism (the most common traits considered in previous works), we also consider: homophobia and transphobia; social class hatred (also referred sometimes as aporophobia); hatred due to physical appearance (e.g., overweight); hatred towards people with disabilities; political hate speech; and hate speech against criminals, prisoners, offenders and other people in conflict with the law. These eight characteristics are listed in Table 1 along with reference names that are used throughout the paper.

IV. CORPUS

This section describes the collection of data, its curation, and annotation procedures in the corpus building process. Our aim was to construct a dataset based on user messages commenting on specific news articles, in a similar fashion to the reader forums present in many news websites. Figure 2 offers a schematic illustration of our dataset, starting with a tweet from a digital newspaper account about China banning the breeding of dogs for human consumption, its respective news article, and replies from users to the original tweet.

A. DATA COLLECTION

Our data collection process started with the official Twitter accounts of a selected set of Argentinian news outlets: La Nación (@lanacion), Clarín (@clarincom), Infobae (@infobae), Perfil (@perfilcom), and Crónica (@cronica). These are the main national newspapers and attract a vast volume of interaction on Twitter. We considered a fixed time period of one year, starting in March 2020. We collected

FIGURE 2. Example of elements in our corpus: a news article (bottom left), a tweet referring to it (top), and Twitter users' replies (bottom right). The user comments are the instances analyzed as potential hate speech; the original tweet and the article itself are the contexts. (All texts in this Figure were translated from Spanish to English).

the replies to each post of the mentioned accounts using the *Spritzer* Twitter API, listening to any tweet mentioning one of their usernames.

For the purpose of this work, we were only interested in the first level of replies to the original tweet, in order to consider as context only the news under debate. If the second or further levels of replies had been considered, the context would have also contained comments made by other users (i.e., a conversational thread), which we wanted to avoid. Also, we discarded tweets from news outlets that were not linked to a news article.

To focus on hate speech related to COVID-19 pandemic, our dataset only kept those articles in which the text body contained at least one of the following terms: coronavirus, COVID-19, COVID, Wuhan, *cuarentena* (quarantine), *normalidad* (normality), *aislamiento* (isolation), *padecimiento* (suffering), *encierro* (confinement), *fase* (phase), *infectado* (infected), *distanciamiento* (distancing), *fiebre* (fever) and *síntoma* (symptom).

Hate speech is not evenly distributed across news articles or topics of discussion. Previous work has focused on multiple strategies to detect users or topics around which this phenomenon is prevalent: for example, monitoring specific targets, hashtags, or offending users [16]. In this case, some form of sampling strategy was also necessary before developing the annotation step, since a random sample of the collected data would have brought a very small quantity of hateful messages.

One of our sampling strategies consisted of using some keywords to select interesting articles, taking into account topics that could be a focus of hate speech. A second strategy sampled articles based on their comments: news containing comments with common insults or pejorative expressions towards the previously defined protected groups. That is, we kept only news articles containing two or more comments that were marked according to a list of predefined insults. We selected expressions and insults that addressed the protected characteristics considered in the hate speech definition, described in Section III. The list of insults and some other technical details are described in Appendix IX-A.

After some trials and subjective evaluation of the articles retrieved using each strategy, we decided to use the latter one — i.e., to select news articles based on their user comments — as it seemed to produce better results. We emphasize that we included in the sample the whole news article and its comments, and not just the replies that contained insults. For each sampled article, 50 comments were randomly chosen for annotation, after excluding those with URLs or images.

Finally, we anonymized tweets by removing user handles and replacing them with a special `@user` token, as some user accounts are usually mentioned by hateful users that could bias the annotation process.

TABLE 2. Annotators profile: gender, age range, education, area of studies. * indicates ongoing. F stands for female, M for male, NB for non-binary.

Gender	Age	Educ.	Area
F	25-30	PhD*	Psychology
NB	31-35	Undergrad	Arts
F	31-35	Undergrad	Anthropology
M	36-40	Graduate	Sociology
F	36-40	PhD	Psychology
F	31-35	Graduate	Communication

B. ANNOTATORS

Considering that hate speech is usually manifested through slang, slurs and insults with a strong socio-cultural background, we hired six Rioplatense Spanish native speakers. That allowed them to be aware of ironic and more subtle forms of hate speech expressions in that variant. As the annotation process considered an initial training step (described in Section IV-C) and in order to label as much data as possible —given the restricted resources— six people was considered a good number in order to have some rotation in the work. Following the lines of Data Statements [61], we provide in this subsection a profile of the annotators.

The recruited annotators were students and/or graduates of social sciences, humanities, or related careers, with no experience in artificial intelligence or data science (to avoid biases). In addition, they were frequent users of social networks so they could capture the subtleties of language in that medium.

As part of the recruitment process, they were asked to take a paid test that consisted in reading the guidelines and annotating ten articles with their respective comments. After this evaluation, no applicants were rejected.

Table 2 provides disaggregated information about the six annotators hired for the task. All six had a high education profile, and two of them had previous experience in data labeling. At the time of the study, two of the annotators were activists in organizations related to some of the vulnerable groups considered in this work. Four of them identified themselves as members of targeted groups: women and LGBTI (lesbian, gay, bisexual, transgender and intersex).

C. ANNOTATION PROCESS

To annotate our data, we followed a similar process to the MAMA portion of the MATTER cycle [62]. First, we defined a model; that is, a practical representation of what we intended to annotate. Figure 3 represents the annotation model used in this work, which follows a hierarchical structure as proposed by [63]. For each comment and its respective context (the tweet from the digital newspaper and linked full article), a first annotation required to mark whether the comment is hateful or not. If it was marked as not hateful, no further information is required. If it was marked as hateful, two extra annotations were required:

- An annotation to indicate whether the comment contains a call to action or not; and
- One or more annotations for each protected characteristic that is attacked in the message.

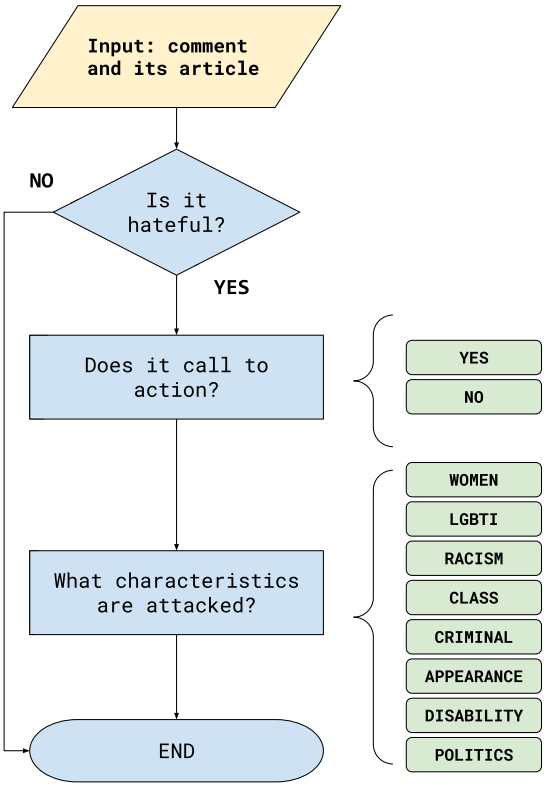


FIGURE 3. Annotation model for each pair of articles and comments.

Each annotation task comprised a news article along with each of the selected comments for it. Annotators were given the option of skipping an article when they considered it irrelevant in terms of hate speech, or when they did not want to annotate it due to personal reasons (no one actually skipped an article due to this).

For each article, up to 50 comments were displayed. The annotator had to label the comments following the hierarchical schema shown in Figure 3. Each article was presented at first to two different annotators with all its comments. Then, a third annotator only had to annotate those comments marked as hateful at least by one of the annotators. While for a majority voting scheme it would just be necessary to check those with exactly one hateful annotation (that is, those comments which have mixed labels), an extra annotation was collected in all cases for further experiments.

Each annotator was required to go through an initial training stage, consisting of the test mentioned in Section IV-B plus the annotation of 15 articles. This set of articles was the only subset labeled by all the annotators and was discarded from the final dataset. At the end of this stage, they were given feedback to adjust their criteria and then proceeded to the actual annotation task.

D. DATASET RESULTS

The resulting dataset consists of 56869 tweets from 1238 news articles. From these tweets, 8715 tweets were marked as hateful by two or three annotators. Table 3 displays the number of hateful tweets for each of the considered

TABLE 3. Number of hateful tweets aggregated by characteristic (i.e. annotated by at least two annotators as hateful), with the respective number of tweets calling to action. Inter-annotator agreement is reported for each characteristic, measured by Krippendorff's alpha.

Characteristic	Count	Calls to action	α
RACISM	2,469	674	0.608
APPEARANCE	1,803	34	0.735
CRIMINAL	1,642	722	0.618
POLITICS	1,428	136	0.509
WOMEN	1,332	18	0.531
CLASS	823	135	0.404
LGBTI	818	11	0.555
DISABLED	580	4	0.596

characteristics. The predominant class of hateful tweets corresponds to racism, followed by tweets offending by appearance.

Calls to action were mainly addressed against criminals and also driven by racist motives. Hateful tweets due to class and political reasons have some calls to action as well, and the other groups did not account for much of these violent expressions. Table 4 displays some examples of hateful tweets with their corresponding annotations.

Among 8715 hateful comments, 77% (6777) contain one attacked characteristic only; nearly 20% have two or more; and 220 comments have three or more. Maximum co-occurrence occurs between the characteristics WOMEN and APPEARANCE, followed by RACISM and CLASS, POLITICS and CLASS, and RACISM and POLITICS. More information about the co-occurrence of attacked characteristics can be found in Appendix IX-B.

As suggested by [40], we checked the distribution of users generating hateful content, so as to avoid having a small number of users responsible for the majority of offensive interactions. Hateful comments per user mean is 1.44, with only 28 users (out of a total of nearly 30,000) having more than ten hateful comments.

Inter-annotator agreement was measured via Krippendorff's alpha [64], using the implementation included in the `krippendorff` library for Python.[9] The agreement for the hate speech label was 0.579, which is compatible with other studies in the area and is expectable considering that we used a rather broad definition of hate speech [17]. For the *calls-to-action* label, the agreement was slightly higher at 0.641. Individual agreements for each characteristic are displayed in Table 3.

To assign gold labels for each tweet in the dataset, we followed a majority-vote strategy. A tweet was marked as hateful if at least two annotators (out of three, at most) labeled it as such. The CALLS label (calls-to-action) was marked if at least two annotators selected it, and we marked each characteristic if at least one annotator selected it. If a tweet was not marked as hateful, no other labels were assigned.

V. CLASSIFICATION EXPERIMENTS

Now with this specially-crafted corpus containing context, we turn our attention to our original research question: can classifiers leverage context to improve their performance in the task of hate speech detection? For this purpose, we proposed the following classification tasks:

- **Binary** hate speech detection: Given a tweet, predict whether it is hateful or not.
- **Fine-grained** hate speech detection: Given a tweet, predict the attacked characteristics (if any), and whether it involves a call to action or not.

In machine learning terms, the binary task can be posed as a binary classification task, while the fine-grained task is a multi-label classification task. Figure 4 illustrates the difference between both tasks as a Venn diagram: in the binary task, we have to predict whether a tweet belongs to the set of hateful tweets; whereas in the fine-grained one, we have to predict if a tweet belongs to the set of hateful tweets for each given characteristic (eight, in our case). The binary task can be seen as a simpler form of the fine-grained task.

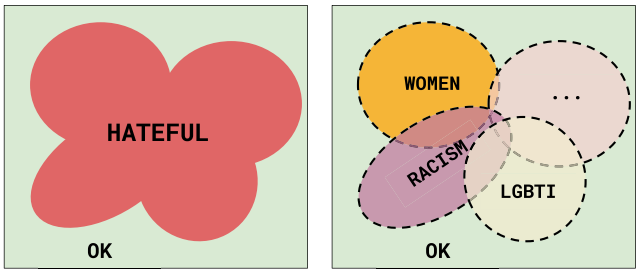


FIGURE 4. Proposed tasks. The binary task consists in predicting whether a tweet is hateful or not. The fine-grained task consists in predicting the attacked characteristics, and whether it calls to action or not.

A. CLASSIFICATION ALGORITHMS

For both tasks, we trained algorithms based on state-of-the-art classification techniques, namely BERT. As explained in Section II, BERT models are based on Transformer language models pre-trained on large corpora. To adapt them for a specific task, a fine-tuning process is performed, which consists in removing the last layer of the language model (usually, a big softmax for the Cloze task[10]), replacing it with a proper layer for the downstream task (e.g. sentiment analysis, question answering), and then adjusting the weights of the whole model [21], [42].

As our dataset is in Spanish, we used BETO [20], a monolingual BERT model for this language. We employed its base version, which consists of 12 Transformer layers with 12 attention heads each, summing up around 100M parameters.

To assess the importance of having contextual information, we considered three different types of inputs for the proposed models: the comment without any context (which we call **None**), the comment with the tweet to which it responds as context (**Tweet**), and the comment with the tweet to which it

[9]https://github.com/pln-fing-udelar/fast-krippendorff

[10]The Cloze task is widely used to evaluate an NLP system's language understanding ability, that consists in replacing a missing part of a text.

TABLE 4. Some hateful examples of our dataset for each of the considered characteristics. NOTE: We translated "bija" — a purposely misspelling of "pija"— as "bick".

Characteristic	Context	Comment
WOMEN	Around the world: Florencia Peña shows her luxurious new house with bar, dock and pool	When you suck the right ones
WOMEN	Mia Khalifa: acted in porn videos for a few months, became world famous and now fights to erase her past	HAHAHAHAHAHAHA KEEP SUCKING....
WOMEN	Narda Lepes: "They touched my ass a thousand times in restaurant kitchens"	Do you have a nice ass?
LGBTI	Why Flor de la V did not continue in Mujeres de eltrece, after the departure of Claudia Fontán	...because she is not a woman, crystal clear
LGBTI	Historical: Mara Gómez was enabled and will be the first trans player in Argentine soccer	What pair of balls this girl has!!!
LGBTI	The story of the Colombian trans model kissing the belly of her eight-month pregnant husband	A male kissing another male
LGBTI	This is what actor Elliot Page looks like after declaring himself trans	she has bick? No. she has pussy? Yes. She is a woman
RACISM	Coronavirus. Yanzhong Huang: "It is quite likely that a Covid-21 is already brewing"	Urgent bombs to that damned race
RACISM	Scientists denounced China's new maneuver to hide the true figures of the coronavirus	Globally we maintain China because everything comes from there and today we are melted and in an emergency... #ChinaVirus I don't want to see a #Chinese for a long time!
RACISM	Impressive operation with tanks for a prosecutor to enter an area controlled by Mapuches	Stop it!!! They are not Mapuches, they are criminals!!! Let's see if someone puts the balls where they have to be put and they shoot them down at once!
CRIMINAL	Rosario: a group of neighbors beats to death a young man accused of stealing cars	this is great, an example to others
CRIMINAL	A guy takes the gun from the thief who assaulted him, runs him off and shoots him dead: arrested	Great, let's go for the total extermination of these apes.
CLASS	Social movements cut off 9 de Julio Av.: they demand a minimum wage of $45,000	get to work, mfs.
POLITICS	A new COVID-19 mutation is confirmed, up to 10 times more contagious than the original strain from Wuhan	I'M VERY GLAD. I HOPE IT WILL ARRIVE SOON IN ARGENTINA AND DESTROY EVERYTHING. WE COULD FINALLY SEE SOMETHING MORE HARMFUL THAN PERONIST CANCER AND ITS KIRCHNERIST METASTASIS.

responds plus the text of the news article (**Full**). The special `[SEP]` token is used to encode the separation between the context and the analyzed text in the **Tweet** and **Full** inputs (our two context-aware models).

For the binary task, we trained a standard BERT architecture for binary sequence classification [21], consisting of a sigmoidal output consuming the last hidden state of the `[CLS]` token, which acts as a continuous representation for the whole sentence. For the fine-grained task, we propose a multi-label output; that is, the simultaneous prediction of the eight characteristics and the call-to-action label. Figure 5 illustrates both models for their three different types of inputs.

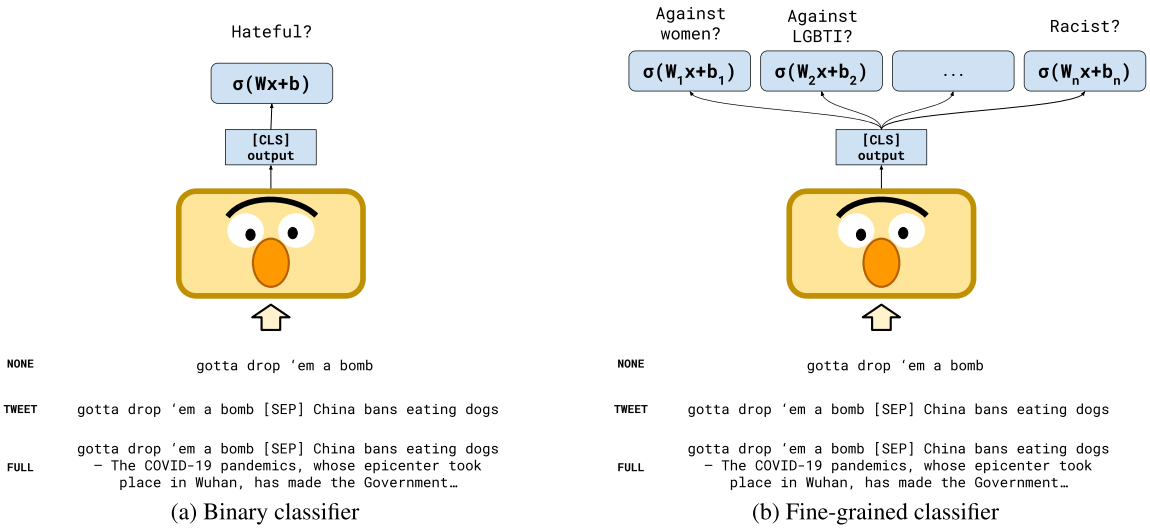
B. TRAINING

We trained the classifiers following the guidelines of [21]. We used Adam [65] as the optimizer, with a weight decay of 0.1, a peak learning rate of $5 * 10^{-5}$ (at the 10% of the optimization steps), and a batch size of 32. We trained the model for 5 epochs, and selected the best model according to the F1 score on the dev set. The loss function for the binary detection task was the binary cross-entropy loss, defined as

$$L_b(y, \widehat{y}) = -y \log(\widehat{y}) - (1 - y) \log(1 - \widehat{y})$$

where y is the true label (0 or 1) and $\widehat{y}$ is the predicted probability of the positive class.

The training process for the fine-grained models was mostly the same, with the exception of the loss function. As the output of the model is a vector of probabilities for each output variable (eight characteristics plus call-to-action), we used a multi-label loss function that considers the probability of each class independently. Let d be the number of output variables (9 in our case), $y \in \{0, 1\}^d$ the true label vector, and $\hat{y} \in [0, 1]^d$ the predicted probabilities. Then, the



(a) Binary classifier (b) Fine-grained classifier

FIGURE 5. Classification models for the proposed tasks. Three different types of classifiers were trained according to the type of input: None (no context), Tweet (context is the tweet to which the comment responds), and Full (context is the tweet to which the comment responds plus the text of the news article).

loss function is defined as:

$$L(y, \widehat{y}) = \sum_{i=1}^{d} L_b(y_i, \widehat{y}_i)$$

where L_b is the binary cross-entropy previously defined.

Sharing the weights between all of the outputs has two benefits: first, it allows for the creation of a more compact model (otherwise there would be nine different BERTs adding up to a billion parameters); and second, it enables sharing common information between the different attacked features. Further details about the training process can be found in Appendix IX-C.

C. DOMAIN ADAPTATION

Standard training of BERT-based classifiers includes two steps as explained in Section V-A: the pre-training of the language model and the fine-tuning of the model to the downstream task [21]. Other transfer-learning approaches in NLP —such as Universal Language Model Fine-tuning [44]— incorporate an intermediate step, that adjusts the pre-trained model to the target domain by continuing the language modeling using the text of the downstream task. Reference [66] showed that continuing the pre-training of BERT-based models on the target domain improves the performance of the models for several subdomains of tasks.

In our experimental setup, we adapted BETO using a sample of comments and articles discarded from the annotation process. As we had three different types of inputs, we performed three domain adaptations according to the shape of the input, as shown in Figure 5.

Table 5 contains the hyperparameters used to adapt the BETO model to our domain. We used the remaining data of the collection process, consisting of around 288000 articles and 5000000 comments. Three versions of BETO were fine-tuned, according to each possible input: no context, tweet, and full context (tweet plus article).

TABLE 5. Hyperparameters used for domain adaptation.

Hyperparameter	Value
Steps	10,000
Batch size	2,048
Max Seq. Length	128, 256 and 512
β_1	0.9
β_2	0.98
ϵ	10^{-6}
Decay	0.01
Peak LR	0.0004
Warmup ratio	0.1

TABLE 6. Results of classification experiments for the binary detection task. Each model is a BETO with three possible inputs: the comment alone without context (None), the comment and the news outlet's tweet (Tweet), and the comment plus the news outlet's tweet plus the article body (Full). Results are expressed as the mean of ten runs of the experiment along with its standard deviation.

	None	Tweet	Full
Precision	71.8 ± 1.6	$\mathbf{74.8 \pm 1.9}$	72.8 ± 2.4
Recall	60.2 ± 1.4	$\mathbf{65.3 \pm 1.4}$	64.1 ± 2.3
F1	65.5 ± 0.4	$\mathbf{69.7 \pm 0.3}$	68.1 ± 0.6
Macro F1	79.8 ± 0.2	$\mathbf{82.2 \pm 0.2}$	81.3 ± 0.3

D. PREPROCESSING

Each tweet was preprocessed using the *pysentimiento* library [67]: we cut character repetitions up to three occurrences; laughs were normalized; user handles were replaced by a special `@user` token; emojis were converted to a text representation. Hashtags were stripped, surrounded by a special `hashtag` token, and segmented to words if they were camel-cased.

TABLE 7. Results of classification experiments for the *fine-grained* task, measured as F1 score for each of the characteristics and macro-averaged metrics. Each model is a BETO with 3 possible inputs: the analyzed comment alone (**None**), the comment plus the tweet from the news outlet (**Tweet**), and comment plus the news outlet's tweet plus the article body (**Full**). Results are expressed as the mean of ten runs of the experiment along with its standard deviation.

	Context		
	None	Tweet	Full
CALLS	65.1 ± 1.9	$\mathbf{68.5 \pm 0.9}$	68.0 ± 1.5
POLITICS	61.1 ± 0.8	62.5 ± 1.3	$\mathbf{64.8 \pm 1.4}$
APPEARANCE	74.2 ± 1.0	$\mathbf{76.6 \pm 0.9}$	75.8 ± 0.9
DISABLED	58.2 ± 1.3	$\mathbf{60.9 \pm 1.8}$	57.8 ± 1.7
WOMEN	38.9 ± 1.5	$\mathbf{42.1 \pm 1.7}$	$\mathbf{42.1 \pm 2.2}$
RACISM	65.3 ± 1.0	$\mathbf{72.0 \pm 0.4}$	71.1 ± 1.0
CLASS	43.3 ± 1.3	$\mathbf{51.1 \pm 2.0}$	47.6 ± 2.7
LGBTI	36.6 ± 1.9	$\mathbf{48.2 \pm 1.9}$	44.5 ± 2.1
CRIMINAL	52.9 ± 1.1	$\mathbf{69.9 \pm 1.9}$	66.8 ± 1.7
Macro F1	55.1 ± 0.5	$\mathbf{61.3 \pm 0.7}$	59.8 ± 0.6
Macro Precision	63.0 ± 1.8	$\mathbf{70.2 \pm 0.9}$	67.8 ± 1.4
Macro Recall	49.9 ± 1.2	$\mathbf{55.1 \pm 1.1}$	54.1 ± 1.3

In order to work with more friendly computational costs, we limited the sequence lengths to 128, 256, and 512 tokens for the **None**, **Tweet** and **Full** model inputs, respectively.

E. EVALUATION

We split our dataset into training, development and test sets to train and evaluate our proposed classifiers. To avoid overestimating the performance, we used a disjoint set of articles for the test set. The training and development splits comprise 36420 and 9120 comments respectively, both coming from 990 articles. The test set has 11343 comments from 248 articles.

Standard metrics were used for both tasks: precision, recall, F1-score and Macro F1 score for the binary classification task. For the fine-grained classification task, we measured F1 for each attacked characteristic, as well as macro-averaged metrics.

VI. RESULTS

Table 6 displays the results of the binary classification task, measured in accuracy, precision, recall, F1, and Macro F1. Results are expressed as the mean of each metric, along with its standard deviation, over ten independent runs of experiments. We present the results only for the domain-adapted BETO classifier; full results can be found in Appendix IX-C. We can observe that the model consuming the simple context (**Tweet**) obtains the best results, with an improvement against the context-unaware (**None**) model of 4.2 F1 points on average. The model with the complete context gets worse results than the model with the simple context, although it improves the general performance against the context-unaware version.

Table 7 shows the results of the classification experiments for the **fine-grained** task, measured by F1 score for each of the features and macro-averaged metrics. As expected, the performance boost of including context is more evident in this task, with a difference of approximately 6 points between the context-unaware and context-aware models (55.1 vs. 61.3 Macro F1). Regarding the two types of context, again the simple version obtains better performance in most of the characteristics, with the only exception of POLITICS.

The characteristics that benefit the most from adding context are CRIMINAL (+17 F1 points), LGBTI (+12), CLASS (+8), and RACISM (almost +7); on the other hand, APPEARANCE and POLITICS benefit the least. It is worth pointing out that, even with the help of added context, some characteristics prove to be very difficult for our classifiers and show a low performance, relatively: WOMEN, LGBTI and CLASS.

A. ERROR ANALYSIS

To have a better understanding of the benefits of adding context and also its limitations, we performed an error analysis between the context-unaware and context-aware models. To do this, we manually checked the output of ten classifiers and looked for their most common errors. Table 8 shows a selection of test instances where context helps to correctly classify comments, and also some examples where both versions are failing to flag them as hateful. We can observe that context helps to disambiguate some of the messages, which are not clearly understood without the additional information.

A remarkable case is that of LGBTI. The mention of any topic-related word in the headline (such as transgender, gay or lesbian) gives some hint to the classifiers about the nature of the message. Nevertheless, due to the complexity of the offenses against transgender individuals (addressing them by their opposite gender, or slurs about their genitals, for instance) models usually fail in flagging these messages as hateful.

VII. DISCUSSION

For the proposed tasks, we could observe that context seems to give a moderate improvement in the binary setting, and

TABLE 8. Error analysis between non-contextualized and contextualized classifiers. Context and comments are shown. The first group of rows (*FN –false negatives– without context, TP –true positives– with context*) represent tweets that were incorrectly labeled as non-hateful by non-contextualized classifiers but contextualized classifiers correctly marked as hateful. The second group consists of tweets that were incorrectly labeled as hateful by non-contextualized classifiers, but contextualized classifiers correctly marked them as non-hateful (FP stands for false positives and TN for true negatives). The last group contains messages that are hateful but were not detected by any classifier, neither non-contextualized nor contextualized.

		Context	Comment
FN without context, TP with context	WOMEN	Ofelia Fernández supported the Government in the controversy over the prisoners and pointed to the Justice that "hates women"	motherfuck*r ,, hopefully you will soon receive a visit from one of those worms. They will fit you. Willing to support him. Government? Fat creeping larva. Debrained
	WOMEN	Did More Rial find love in a personal trainer?	You have to be hungry to eat that bolivian piglet
	LGBTI	What Elliot Page looks like after declaring himself transgender	hope she gets psychiatric help
	LGBTI	Mara Gómez fulfills her dream: she will be the first transgender footballer in the Argentine professional tournament	Mara "the club" Gómez
	LGBTI	Mara Gómez fulfills her dream: she will be the first transgender footballer in the Argentine professional tournament	go break some legs boy
FP without context, TN with context	LGBTI	A man got into his car at the door of the Chinese Embassy and claimed that he had explosives	He is not a man. He's a jerk
	LGBTI	Coronavirus in Argentina: 70% of cases are in men	The corona is female
	LGBTI	The ruling party calls for a "federal caravan" in support of the Government and the tax on large fortunes	Gross
	CLASS	Paul McCartney: "The Chinese need to be cleaner and less medieval"	it had to be said at last
	CLASS	Main teaching union rejected the return to presential classes	shitty bums!
Not detected by any classifier	WOMEN	Why Women-Led Countries Appear To Have Responded Better To The Coronavirus Crisis	because they wash, iron and sweep?
	WOMEN	British girl went to Peru for 10 days and stayed for love: she lives with no water and among insects	she left everything coz of the wood of that Peruvian ahaha that nigger must have a generous dick
	WOMEN	Did More Rial find love in a personal trainer?	gotta be well-trained to lift that hippo
	CLASS	The Government will spend $75B to develop 300 slums throughout the country	without education behind this is nothing, they will remain the same old misfits but now with Netflix.
	LGBTI	She told that she was a lesbian, her father confessed that he was gay and now his mother fell in love with a woman: this is how he was inspired for his second film	The film is called the failure of a normal family
	LGBTI	"Why don't we see trans doctors?": The claim of a prestigious cardiologist for America to be more inclusive	because sick people cannot heal sick people
	LGBTI	A trans woman is killed in Rosario after a burst of 20 shots	Why did she not pull out her shotgun and apply self-defense?!

a more considerable gain in the fine-grained setting. This result might appear to contradict recent work that found no improvement by means of contextualization in toxicity detection [55]. However, it must be noted that hate speech is one of the most complex forms of toxic behavior; thus, hate speech detection might benefit differently from having additional information. Also, while [55]'s context was extracted from the entire conversation preceding the target message, our context was taken from the news' tweet and the article itself under discussion. Further, [57] recently found that toxicity detection algorithms can take advantage of additional information by restricting the analysis to a subset of context-sensitive comments.

Something interesting this dataset provides is a characterization of hate speech. Since we have the attacked characteristics for each hateful tweet, we could assess the influence of context for each protected characteristic. Contextual information seems to have more impact on some characteristics than others (e.g., when the attack is against LGBTI people). Moreover, we can observe that the dataset has complex and compositional examples of discriminatory language for specific characteristics.

The constructed dataset has both short and long contexts. In our experiments, we have observed no substantial improvement in model performance by using the long context; that is, the full article. This might coincide with a familiar behavior observed in humans —that many people comment after reading nothing but the headline. (However, it might be argued that humans have access to a richer context and information beyond the headline.)

The experiments performed in this work have a few limitations. First, human annotators had access to the full contexts when doing their task. To better assess the impact of context in hate speech detection, context-unaware models should be trained on comments labeled by humans without access to any additional information. Second, a practical limitation is that context is not always available for any given text. Even if it was able to find one, it might not always consist of a news article — it may also be a conversational thread, or even audiovisual content, for example. Lastly, the labeled comments are replies to tweets published by media outlets, which limits the possible forms of our instances. Therefore, further study is needed to understand how other forms of messages and contexts impact the detection of hate speech.

VIII. CONCLUSION

In this work, we have assessed the impact of adding context to the automatic detection of hate speech. To do this, we built a dataset consisting of user replies to posts on Twitter published by main news outlets in Argentina, and annotated it using carefully designed guidelines. We conducted a series of classification experiments using transformer-based techniques, and found clear evidence that certain contextual information leads to improved performance: our models showed a 4 to 5 point increase in Macro F1 after adding context.

Although in our experiments the smallest context (the news article tweet) was the one that obtained the best results, a future line of work could explore ways to include other sources of information. For instance, adding real-world knowledge about the targets of hate speech could be useful. This information might be even available in the news article itself, or other sources such as a knowledge graph.

From the perspective of error analysis, it can be seen that some categories of hate speech are elusive for state-of-the-art detection algorithms. One of these cases is the abusive messages against the LGBTI community, which contain semantically complex messages, with ironic content and metaphors that are difficult to interpret for classifiers based on state-of-the-art language models. Despite these limitations, the detection of hate speech against the LGBTI community was among the most benefited by the addition of context. Future work should explore the reasons behind the difficulties for the state-of-the-art models to detect it, and also explore ways to improve the detection of this type of hate speech.

We may conclude that hate speech detection clearly benefits from the use of **contextual information**. The evidence from our experiments —preliminary for now, and with the limitations noted in the discussion— indicates that state-of-the-art models can use this information to improve the detection of hate speech in social networks. We hope that this work will encourage the use of contextual information in the detection of hate speech and other opinion-mining tasks and that it will be a starting point for future research in this area.

TABLE 9. Seed expressions used to select articles based on possibly hateful comments.

Expression	Description or translation
viejo puto	old fag
marica	fag
sodomita	sodomite
degenerados	degenerate
trabuco, trava	slur for transgender woman
travesti	transgender woman
bija	misspelling of dick
feministas	feminists
feminazis	offensive term against feminists
aborteras	abortion activists
gorda	fat woman
uno menos	one less (celebratory expression for a killing)
urraca	magpie (offensive slur against a woman)
prostituta	prostitute
putita	little bitch
reventada	prostitute
peruano, peruca	peruvian
paraguayo	paraguayan
trolo	fag
bala	bullet (as in "shoot them"); also fag
bolita	slur for bolivian
negro(s) (de)	nigger
judío, sionista	jew, zionist
matarlos	(have to) kill them
chinos	chinese
una bomba	a bomb
vayan a laburar/trabajar	go to work
villeros	shanty dwellers

TABLE 10. Number of articles and comments in the dataset per news outlet.

Newspaper	#Art	#Comm
@infobae	590	26,834
@clarincom	370	17,501
@LANACION	222	10,378
@cronica	42	1,562
@perfilcom	14	594
Total	1,238	56,869

AVAILABILITY OF DATA AND MATERIAL

We have made our corpus available at the huggingface hub.[11] For the sake of reproducibility and also for further research, we will release the anonymized annotations (as suggested by [68]) in addition to the aggregated dataset, as well as annotation guidelines.

SUPPLEMENTAL MATERIAL

A. DATA SELECTION

Table 9 lists the seed expressions used to mark potentially hateful comments. This list was constructed manually, checking for some common expressions in the data. We used MongoDB's text index to retrieve any comments containing at least one of them.

Some of these expressions were used literally (with quotation marks) and some were allowed inflections provided by the search engine. For some of them, we excluded other words: for instance, when querying "negra" (*female nigger*) we removed "plata | guita" (*money*) as there were many hits for such queries. For others, we added prepositions to the query (such as "negro de") because using just "negro" had a lot of non-hateful hits.

[11] https://huggingface.co/datasets/piuba-bigdata/contextualized_hate_speech

(a) Co-occurrence of attacked characteristics within the same comment

(b) Co-occurrence of attacked characteristics within the same article

FIGURE 6. Co-occurrence matrices for attacked characteristics in hateful messages. Figure 6 a shows co-occurrence within the same comment, and Figure 6b shows co-occurrence across comments of the same article. Brighter indicates more co-occurrence.

TABLE 11. Results of the classifiers for the binary task, expressed as the mean and standard deviation of ten independent runs of the experiments. Three different types of inputs are considered: no context, comment + tweet of the news outlet, and full context. FT means that the pre-trained language model was fine-tuned, and ¬FT means it was not fine-tuned.

	Context					
	None		Tweet		Full	
Metric	BETO	BETO$_{FT}$	BETO	BETO$_{FT}$	BETO	BETO$_{FT}$
Accuracy	88.9 ± 0.3	89.9 ± 0.2	90.2 ± 0.2	91.0 ± 0.2	90.4 ± 0.2	90.5 ± 0.3
Precision	67.8 ± 2.0	71.8 ± 1.6	73.1 ± 1.1	74.8 ± 1.9	73.9 ± 1.6	72.8 ± 2.4
Recall	56.8 ± 1.7	60.2 ± 1.4	60.1 ± 1.0	65.3 ± 1.4	61.1 ± 1.6	64.1 ± 2.3
F1	61.8 ± 0.5	65.5 ± 0.4	66.0 ± 0.6	69.7 ± 0.3	66.9 ± 0.5	68.1 ± 0.6
Macro F1	77.6 ± 0.3	79.8 ± 0.2	80.1 ± 0.3	82.2 ± 0.2	80.6 ± 0.2	81.3 ± 0.3

TABLE 12. Results of the classifiers for the fine-grained task, expressed as the mean and standard deviation of ten independent runs of the experiments. Three different types of inputs are considered: no context, comment + tweet of the news outlet, and full context. FT means that the pre-trained language model was fine-tuned, and ¬FT means it was not fine-tuned.

	Context					
	None		Tweet		Full	
Metric	¬FT	FT	¬FT	FT	¬ FT	FT
CALLS	64.6 ± 1.0	65.1 ± 1.9	63.8 ± 0.9	68.5 ± 0.9	65.3 ± 1.3	68.0 ± 1.5
WOMEN	37.3 ± 1.3	38.9 ± 1.5	41.1 ± 0.9	42.1 ± 1.7	38.1 ± 1.7	42.1 ± 2.2
LGBTI	35.1 ± 1.8	36.6 ± 1.9	45.1 ± 2.1	48.2 ± 1.9	42.7 ± 2.4	44.5 ± 2.1
RACISM	63.5 ± 1.4	65.3 ± 1.0	68.8 ± 1.2	72.0 ± 0.4	69.1 ± 0.9	71.1 ± 1.0
CLASS	40.1 ± 1.6	43.3 ± 1.3	49.1 ± 2.2	51.1 ± 2.0	45.1 ± 1.9	47.6 ± 2.7
POLITICS	55.5 ± 1.8	61.1 ± 0.8	57.9 ± 1.4	62.5 ± 1.3	59.1 ± 1.3	64.8 ± 1.4
DISABLED	55.1 ± 1.6	58.2 ± 1.3	58.5 ± 1.6	60.9 ± 1.8	55.7 ± 2.3	57.8 ± 1.7
APPEARANCE	72.6 ± 1.0	74.2 ± 1.0	74.1 ± 1.2	76.6 ± 0.9	75.5 ± 0.9	75.8 ± 0.9
CRIMINAL	51.3 ± 1.4	52.9 ± 1.1	65.0 ± 1.2	69.9 ± 1.9	65.4 ± 2.3	66.8 ± 1.7
Macro Precision	55.8 ± 1.0	63.0 ± 1.8	64.2 ± 1.6	70.2 ± 0.9	67.7 ± 1.4	67.8 ± 1.4
Macro Recall	50.6 ± 0.6	49.9 ± 1.2	54.0 ± 0.8	55.1 ± 1.1	50.4 ± 0.9	54.1 ± 1.3
Macro F1	52.8 ± 0.5	55.1 ± 0.5	58.2 ± 0.5	61.3 ± 0.7	57.3 ± 0.7	59.8 ± 0.6

It is important to stress that this method was only used for selecting news articles for the subsequent annotation step, and comments were randomly sampled among the replies to the selected articles.

B. ADDITIONAL INFORMATION of the DATASET

Table 10 displays the number of articles and comments in the final dataset. We can observe that most articles and comments come from @infobae, followed by @clarincom and @LANACION.

From the 8715 hateful comments present in the dataset, 77% of them (6777) contain only one attacked characteristic, nearly 20% have exactly two, and 220 comments have three or more. Figure 6 illustrates the co-occurrence matrix between the different characteristics for comments having more than one attacked characteristic. We can observe that the maximum co-occurrence occurs between the characteristics WOMEN and APPEARANCE, followed by RACISM and CLASS, POLITICS and CLASS, and RACISM and POLITICS.

Another way of analyzing co-occurrence is by grouping the different characteristics of their comments by articles, to observe how the same context can invoke different types of discrimination. Figure 6a illustrates the interactions between the different characteristics per article. Greater dispersion is observed in the co-occurrences than in Figure 6b, showing some additional interactions such as between RACISM and POLITICS and —perhaps unexpectedly— between APPEARANCE and POLITICS.

C. CLASSIFICATION EXPERIMENTS

Table 11 and Table 12 display the full results for the binary and fine-grained tasks. We used two pre-trained language models as our base models: BETO, without any fine-tuning on the data (marked as ¬FT), and a BETO fine-tuned with the remaining data of the collection process, as described in Section V-A. The results show that, in all cases, the fine-tuning process improves the performance of the classifiers.

To train our classification models, we used the *HuggingFace* library [69] and the *PyTorch* framework [70]. We used a *NVIDIA GeFORCE GTX 1080 Ti* to fine-tune the models. To perform the domain-adaptation of the language models, we used a *TPU v2-8* in a *Google Colab Pro* instance, taking 10 hours at its maximum sequence length.

ACKNOWLEDGMENT

The authors would like to thank the annotators who worked to ensure the accuracy and quality of the data used in this study. Their dedication and hard work were essential to the success of this project. They would also like to thank Dr. Eugenia Mitchelstein, who provided valuable insights and suggestions that helped shape the direction of this research. The authors would like to thank CONICET and Universidad Torcuato Di Tella for their support.

REFERENCES

[1] *Hate Speech Explained: A Toolkit*, vol. 19, London, U.K., 2015.
[2] K. Saha, E. Chandrasekharan, and M. De Choudhury, "Prevalence and psychological effects of hateful speech in online college communities," in *Proc. 10th ACM Conf. Web Sci.*, Jun. 2019, pp. 255–264.
[3] M. Bilewicz and W. Soral, "Hate speech Epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization," *Political Psychol.*, vol. 41, no. S1, pp. 3–33, Aug. 2020.
[4] E. Blout and P. Burkart, "White supremacist terrorism in charlottesville: Reconstructing unite the Right," *Stud. Conflict Terrorism*, pp. 1–22, Jan. 2021.
[5] R. McIlroy-Young and A. Anderson, "From 'welcome new gabbers' to the Pittsburgh synagogue shooting: The evolution of gab," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 13, 2019, pp. 651–654.
[6] A. Warofka, "An independent assessment of the human rights impact of Facebook in Myanmar," Facebook Newsroom, vol. 5, Nov. 2018.
[7] T. H. Paing, "Zuckerberg urged to take genuine steps to stop use of Fb to spread hate in Myanmar," Irrawaddy.
[8] European Union. (2016). *The Eu Code of Conduct on Countering Illegal Hate Speech Online*. [Online]. Available: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en
[9] *United Nations Guidance Note on Addressing and Countering COVID-19 Related Hate Speech*, United Nations, New York, NY, USA, 2020.
[10] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, "The proximal origin of SARS-CoV-2," *Nature Med.*, vol. 26, no. 4, pp. 450–452, Apr. 2020.
[11] J. Cohen, "Scientists 'strongly condemn' rumors and conspiracy theories about origin of coronavirus outbreak," Tech. Rep., 2020.
[12] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.
[13] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proc. Int. AAAI Conf. Web Social Media*, vol. 11, no. 1, pp. 512–515, May 2017.
[14] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*, 2017, pp. 1–10.
[15] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, 2018.
[16] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.* Vancouver, BC, Canada: Association for Computational Linguistics, 2019, pp. 54–63.
[17] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: A systematic review," *Lang. Resour. Eval.*, vol. 55, no. 2, pp. 477–523, Jun. 2021.
[18] M. E. Aragón, M. A. A. Carmona, M. Montes-y Gómez, H. J. Escalante, L. V. Pineda, and D. Moctezuma, "Overview of MEX-A3T at iberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets," in *Proc. Iberian Lang. Eval. Forum (IberLEF)*, 2019, pp. 478–494.
[19] E. Fersini, M. Anzovino, and P. Rosso, "Overview of the task on automatic misogyny identification at IberEval," in *Proc. 3rd Workshop Eval. Human Lang. Technol. Iberian Lang. (IberEval) Co-Located 34th Conf. Spanish Soc. Natural Lang. Process. (SEPLN)*, Seville, Spain, 2018, pp. 1–15.
[20] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained BERT model and evaluation data," in *Proc. ICLR*, 2020, pp. 1–10.
[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423
[22] N. Torres and V. Taricco, "Los discursos de odio como amenaza a los derechos humanos," CELE, Tech. Rep., 2019.
[23] *Discurso de Odio y La Incitación a La Violencia Contra Las Personas Lesbianas, Gays, Bisexuales, Trans e Intersex en América*, Comisión Interamericana sobre Derechos Humanos, Tech. Rep., 2015.
[24] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," in *Proc. 1st Workshop Abusive Lang. Online*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 78–84. [Online]. Available: https://aclanthology.org/W17-3012
[25] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*. 25th ed. Dallas, TX, USA: SIL International, 2022.
[26] M. A. A. Carmona, E. Guzmán-Falcón, M. M. Y. Gómez, H. J. Escalante, L. V. Pineda, V. Reyes-Meza, and A. R. Sulayes, "Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets," in *Proc. 3rd Workshop Eval. Human Lang. Technol. Iberian Lang. (IberEval), Co-Located 34th Conf. Spanish Soc. Natural Lang. Process. (SEPLN)*, Seville, Spain, 2018, pp. 1–23.
[27] M. E. Aragón, M. A. A. Carmona, M. Montes-y-Gómez, H. J. Escalante, L. V. Pineda, and D. Moctezuma, "Overview of MEX-A3T at IberLef 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets," in *Proc. Iberian Lang. Eval. Forum Co-Located 35th Conf. Spanish Soc. Natural Lang. Process.*, Bilbao, Spain, M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, J. Carrillo-de-Albornoz, S. Montalvo, L. Chiruzzo, S. Collovini, Y. Gutiérrez, S. M. J. Zafra, M. Krallinger, M. Montes-y-Gómez, R. Ortega-Bueno, and A. Rosá, Eds., vol. 2421. Sep. 2019, pp. 478–494. [Online]. Available: http://ceur-ws.org/Vol-2421/MEX-A3T_overview.pdf
[28] Y. Hswen, X. Xu, A. Hing, J. B. Hawkins, J. S. Brownstein, and G. C. Gee, "Association of '# COVID19'versus '# Chinesevirus' with anti-asian sentiments on Twitter: March 9–23, 2020," *Amer. J. Public Health*, vol. 111, no. 5, pp. 956–964, 2021.

[29] J. Uyheng and K. M. Carley, "Characterizing network dynamics of online hate communities around the COVID-19 pandemic," *Appl. Netw. Sci.*, vol. 6, no. 1, pp. 1–21, Mar. 2021.

[30] M. Li, S. Liao, E. Okpala, M. Tong, M. Costello, L. Cheng, H. Hu, and F. Luo, "COVID-HateBERT: A pre-trained language model for covid-19 related hate speech detection," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 233–238.

[31] *Hidden for Anonimity Requirements*, Anonymous, 2020.

[32] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2004, pp. 468–469.

[33] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proc. 2nd Workshop Lang. Social Media*. Montréal, QC, Canada: Association for Computational Linguistics, 2012, pp. 19–26.

[34] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 85–90. [Online]. Available: http://aclweb.org/anthology/W17-3013

[35] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on Twitter," in *Proc. 1st Workshop Abusive Lang. Online*. Toronto, ON, Canada: Association for Computational Linguistics, 2017, pp. 41–45. [Online]. Available: http://aclweb.org/anthology/W17-3006

[36] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, 2017, pp. 759–760.

[37] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. Adv. Inf. Retr. 40th Eur. Conf. IR Res. (ECIR)*, Grenoble, France, Mar. 2018, pp. 141–153, doi: 10.1007/978-3-319-76941-7_11.

[38] A. Bisht, "Detection of hate speech and offensive language in Twitter data using LSTM model," in *Recent Trends in Image and Signal Processing in Computer Vision*. Cham, Switzerland: Springer, 2020, pp. 243–264.

[39] J. M. Pérez and F. M. Luque, "Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 64–69.

[40] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2019, pp. 45–54.

[41] F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," *Exp. Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114120.

[42] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Tech. Rep., 2018.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[44] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*. Melbourne, FL, Australia: Association for Computational Linguistics, vol. 1, Jul. 2018, pp. 328–339. [Online]. Available: https://aclanthology.org/P18-1031

[45] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III, "A neural network for Factoid question answering over paragraphs," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 633–644.

[46] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[47] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[48] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*. Vancouver, BC, Canada: Association for Computational Linguistics, Nov. 2020, pp. 2898–2904. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.261

[49] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*. Vancouver, BC, Canada: Association for Computational Linguistics, Oct. 2020, pp. 9–14. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.2

[50] J. D. L. Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury, "BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling," *Procesamiento Del Lenguaje Natural*, vol. 68, pp. 13–23, Mar. 2022. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403

[51] A. G. Fandi no, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas, "Maria: Spanish language models," *Procesamiento Del Lenguaje Natural*, vol. 68, 2022.

[52] J. M. Pérez, D. A. Furman, L. Alonso Alemany, and F. M. Luque, "Robertuito: A pre-trained language model for social media text in Spanish," in *Proc. Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 7235–7243. [Online]. Available: https://aclanthology.org/2022.lrec-1.785

[53] D. Nozza, F. Bianchi, and D. Hovy, "What the [MASK]? Making sense of language-specific BERT models," 2020, *arXiv:2003.02912*.

[54] L. Gao and R. Huang, "Detecting online hate speech using context aware models," Sep. 2017, pp. 260–266, doi: 10.26615/978-954-452-049-6_036.

[55] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos, "Toxicity detection: Does context really matter?" in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4296–4305.

[56] H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on Arabic social media," in *Proc. 1st Workshop Abusive Lang. Online*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 52–56. [Online]. Available: https://aclanthology.org/W17-3008

[57] A. Xenos, J. Pavlopoulos, and I. Androutsopoulos, "Context sensitivity estimation in toxicity detection," in *Proc. 5th Workshop Online Abuse Harms (WOAH)*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2021, pp. 140–145. [Online]. Available: https://aclanthology.org/2021.woah-1.15

[58] A. Sheth, V. L. Shalin, and U. Kursuncu, "Defining and detecting toxicity on social media: Context and knowledge are key," *Neurocomputing*, vol. 490, pp. 312–318, Jun. 2022.

[59] M. Wiegand, J. Ruppenhofer, and E. Eder, "Implicitly abusive language—What does it actually look like and why are we not getting there?" in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 576–587.

[60] *General Comment no. 20: Non-Discrimination in Economic Social and Cultural Rights*, UNC Economic Social C Rights, 2009.

[61] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," in *Proc. Trans. Assoc. Comput. Linguistics*, vol. 6, 2018, pp. 587–604. [Online]. Available: https://aclanthology.org/Q18-1041

[62] J. Pustejovsky and A. Stubbs, *Natural Language Annotation for Machine Learning: A Guide to Corpusbuilding for Applications*. Sebastopol, CA, USA: O'Reilly Media, 2012.

[63] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Minneapolis, MN, USA, Jun. 2019, pp. 1415–1420. [Online]. Available: https://aclanthology.org/N19-1144

[64] K. Krippendorff, "Computing krippendorff's alpha-reliability," Tech. Rep., 2011.

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[66] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 8342–8360. [Online]. Available: https://aclanthology.org/2020.acl-main.740

[67] J. Manuel Pérez, J. Carlos Giudici, and F. Luque, "Pysentimiento: A Python toolkit for sentiment analysis and SocialNLP tasks," 2021, *arXiv:2106.09462*.

[68] V. Basile, "It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks," in *Proc. AIxIA Discuss. Papers Workshop*, vol. 2776, 2020, pp. 31–40.

[69] T. Wolf, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.

[70] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, and G. Chanan, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

JUAN MANUEL PÉREZ received the M.Sc. and Ph.D. degrees in computer science from Universidad de Buenos Aires, in 2016 and 2022, respectively. He is currently a Postdoctoral Researcher with Instituto de Ciencias de la Computación, Facultad de Ciencias Exactas, UBA/CONICET, where he is working on contextualized detection of hate speech and opinion mining tasks in social networks using NLP techniques.

FRANCO M. LUQUE received the degree and Ph.D. degree in computer science from the Faculty of Mathematics, Astronomy, Physics and Computing (FAMAF), National University of Córdoba, Argentina. He is an Adjunct Professor with FAMAF, National University of Córdoba, and an Assistant Researcher with CONICET. His main area of research is NLP. He is also involved in research projects on various topics, such as the analysis of hate speech in social networks, multimodal learning for visual dialog, and information extraction in medical reports.

DEMIAN ZAYAT received the master's degree from Stanford University, in 2009, and the Law degree from Universidad de Buenos Aires, in 2000. He is a Constitutional Law Professor. His research interests include human rights and discrimination and combining legal and data-driven research.

MARTÍN KONDRATZKY received the degree in linguistics from Universidad de Buenos Aires and the master's degree in statistics. He is currently an NLP Engineer in the field of information retrieval and question-answering.

AGUSTÍN MORO received the Ph.D. degree in sociology from Universidad Nacional del Centro. He is a Professor of scientific research methodology with Universidad Nacional del Centro. He is also a consultant in the field of evidence-based policy implementation processes.

PABLO SANTIAGO SERRATI is currently pursuing the Ph.D. degree in social sciences with the Faculty of Social Sciences, Universidad de Buenos Aires (UBA). He is currently working on the use of data analysis techniques and statistical programming in the fields of urban studies and social science.

JOAQUÍN ZAJAC received the degree in sociology from Universidad de Buenos Aires, the master's degree in social anthropology, and the Ph.D. degree in social sciences. He is a Postdoctoral Fellow with CONICET, Universidad de San Martin. He specializes in violence, human rights, and security issues, combining qualitative and quantitative data analysis for his research work.

PAULA MIGUEL received the degree and Ph.D. degree in social sciences. She is a Professor of sociology and a CONICET Researcher with Universidad de Buenos Aires, specializing in cultural analysis, qualitative analysis, data science, and NLP techniques.

NATALIA DEBANDI received the degree in computer science from the University of Buenos Aires (UBA), and the Ph.D. degree in social sciences from UBA and University Paris IV Sorbonne. She specializes in the design of human rights indicators and data analysis for academic and applied human rights research.

AGUSTÍN GRAVANO received the Ph.D. degree in computer science from Columbia University, in 2009. He is an Associate Professor with Universidad Torcuato Di Tella and an Independent Researcher with CONICET, Argentina. His research interest includes building computational models of coordination in spoken dialogue, for later improving the naturalness of spoken dialogue systems.

VIVIANA COTIK received the Ph.D. degree in computer science, in 2018. She is an Assistant Professor with Universidad de Buenos Aires and a Research Assistant with CONICET, Argentina. Her research interests include NLP and data science, and she focuses on information extraction from Spanish biomedical texts. She also works on the detection of epidemics from texts.