The Vulnerable Identities Recognition Corpus (VIRC) for Hate Speech Analysis

Ibai Guillén-Pacho[1,*,†], Arianna Longo[2,3,†], Marco Antonio Stranisci[2,3], Viviana Patti[2] and Carlos Badenes-Olmedo[1,4]

[1]Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[2]University of Turin, Italy
[3]Aequa-tech, Torino, Italy (aequa-tech.com)
[4]Computer Science Department, Universidad Poltécnica de Madrid, Spain

Abstract

This paper presents the Vulnerable Identities Recognition Corpus (VIRC), a novel resource designed to enhance hate speech analysis in Italian and Spanish news headlines. VIRC comprises 880 headlines, manually annotated for vulnerable identities, dangerous discourse, derogatory expressions, and entities. Our experiments reveal that recent large language models (LLMs) struggle with the fine-grained identification of these elements, underscoring the complexity of detecting hate speech. VIRC stands out as the first resource of its kind in these languages, offering a richer annotation scheme compared to existing corpora. The insights derived from VIRC can inform the development of sophisticated detection tools and the creation of policies and regulations to combat hate speech on social media, promoting a safer online environment. Future work will focus on expanding the corpus and refining annotation guidelines to further enhance its comprehensiveness and reliability.

Keywords

hate speech, vulnerable identities, annotated corpora

1. Introduction

Hate Speech (HS) detection is a task with a high social impact. Developing technologies that are able to recognize these forms of discrimination is not only crucial to enforce existing laws but it also supports important tasks like the moderation of social media contents. However, recognizing HS is challenging. Verbal discrimination takes different forms and involves a number of correlated phenomena that make difficult to reduce HS as a binary classification.

Analyzing the recent history of corpora annotated for HS it is possible to observe the shift from very broad categorizations of hatred contents to increasingly detailed annotation schemes aimed at understanding the complexity of this phenomenon. High-level schemes including dimensions like "hateful/offensiveness" [1] or "sexism/racism" [2] paved the way for more sophisticated attempts to formalize such concepts in different directions: exploring the interaction between HS and vulnerable targets [3, 4, 5]; studying the impact of subjectivity [6, 7]; identifying the triggers of HS in texts [8, 9].

Despite this trend, the complex semantics of HS in texts is far from being fully explored. Information Extraction (IE) approaches to HS annotation have been rarely implemented, yet. Therefore, corpora that includes fine-grained structured semantic representation of HS incidents are not available. The only notable exception is the recent work of Büyükdemirci et al. [10], which treat the identification of HS targets as a span-based task.

In order to fill this gap, we present the Vulnerable Identities Recognition Corpus (VIRC): a dataset of 880 Italian and Spanish headlines against migrants aimed at providing an event-centric representation of HS against vulnerable groups. The annotation scheme is built on four elements:

- **Named Entity Recognition (NER)**. All the named entities that are involved in a HS expression: 'location', 'organization', and 'person'.
- **Vulnerable Identity mentions**. Generic mentions related to identities target of HS as they are defined by the international regulatory frameworks [1]: 'women', 'LGBTQI', 'ethnic minority', and 'migrant'.
- **Derogatory mentions**. All mentions that negatively portray people belonging to vulnerable groups.
- **Dangerous speech**. The part of the message that is perceived as hateful against named entities or vulnerable identities.

In this paper we present a preliminary annotation experiment intended to validate the scheme and to assess the impact on disagreement in such a fine-grained task. The paper is structured as follows. In Section 2, we discuss related work, in Section 3, we describe the methodology used, in Section 4, we introduce the VIRC corpus, and in Section 5, we present the conclusions and discuss possible future work.

2. Related Work

Literature on automatic HS detection is vast and follows different research directions [11]: from the analysis of subjectivity in the perception of this phenomenon [12] to the definition of ever more refined categorizations of hateful contents [13]. In this section we focus on the approaches to HS detection that are aimed at studying the target of HS inspired by Information Extraction (IE) approaches. In Section 2.1 we review HS

[1]https://www.coe.int/en/web/combating-hate-speech/recommendation-on-combating-hate-speech

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.
†These authors contributed equally.
ibai.guillen@upm.es (I. Guillén-Pacho); arianna.longo401@edu.unito.it (A. Longo); marcoantonio.stranisci@unito.it (M. A. Stranisci); viviana.patti@unito.it (V. Patti); carlos.badenes@upm.es (C. Badenes-Olmedo)
https://iguillenp.github.io/ (I. Guillén-Pacho); https://marcostranisci.github.io/ (M. A. Stranisci); https://www.unito.it/persone/vpatti (V. Patti); https://about.me/cbadenes (C. Badenes-Olmedo)
0000-0001-7801-8815 (I. Guillén-Pacho); 0009-0005-8500-1946 (A. Longo); 0000-0001-9337-7250 (M. A. Stranisci); 0000-0001-5991-370X (V. Patti); 0000-0002-2753-9917 (C. Badenes-Olmedo)

resources inspired by this approach with a specific focus on span-based annotated corpora. In Section 2.2 we discuss the implementation of NER-based techniquest in the creation of HS corpora.

2.1. Hate Speech Detection

A large amount of work on HS detection focuses on classification, both binary (existence or not) and multi-labeled (misogyny, racism, xenophobia, etc.). This has led to the existence of large collections of datasets such as those grouped by [14]. One of the main problems is that most resources are in English, and for mid-to-low resource languages (e.g., Italian), some HS categories are not covered. This constraint is mitigated by cross-lingual transfer learning to exploit resources in other languages [15] and, although good results are achieved, the creation of resources for these languages is still necessary.

The main resources for the identification of HS are particularly focused on a target by identifying the presence or absence of HS in them. As in the work of [16], where in 1,100 tweets in Italian with special target on immigrants were annotated according to the presence of HS, irony, and the stance of the message's author on immigration matters. However, recently, there has been an increasing focus on identifying hateful expressions and their intended targets. The change in paradigm suggests that resources should be wider in scope and not focus on a particular discourse target. The main resources in this field have high linguistic diversity, although they do not all follow the same annotation scheme, with English being the most common language. We have found works in English [17]; Vietnamese[18]; Korean [19]; English and Turkish [10]; and English, French, and Arabic [20]. However, we have not found any in Italian or Spanish, which we believe makes this work the first to cover these languages for this task.

Two main annotation approaches can be drawn from these studies, those that annotate at the span level [17, 18, 19, 10] and those that annotate over the full text [20]. On the one hand, the work that follows the latter approach presents a corpus of 13.000 tweets (5.647 English, 4.014 French, and 3.353 Arabic) and notes the sentiment of the annotator (shock, sadness, disgust, etc.), hostility type (abusive, hateful, offensive, etc.), directness (direct or indirect), target attribute (gender, religion, disabled, etc.) and target group (individual, women, African, etc.).

On the other hand, works that follow the approach of span annotation design different annotation criteria. The simplest, [17, 18], only annotates one dimension. The first, [17], annotates the parts that make a comment toxic on a 30.000 English comments of the Civil Comments platform. The second, [18], annotates only the parts that make a comment offensive or hateful in 11.000 Vietnamese comments on Facebook and Youtube. The other papers, [19, 10], extend this approach and also label the span in which the target of the attack is mentioned. Moreover, [19] is not limited to that; they also annotate the target type (individual, group, other), the target attribute (gender, race, ethnic, etc.) and the target group (LGBTQ +, Muslims, feminists, etc.). Their final corpus has 20.130 annotated offensive Korean-language news and video comments.

However, the guidelines used by the different works sometimes present incompatibilities. Although some works use offensive and hateful labels in the same way [19, 18], others distinguish between these two types of expression [10]. This resource, the last one, has separately annotated hateful and offensive expressions, totaling 765 tweets in English and 765 tweets in Turkish.

2.2. Named Entity Recognition

Developed as a branch of Information Extraction (IE), Named Entity Recognition (NER) is a field of research aimed at detecting named entities in documents according to different schemes. Following the review of Jehangir et al. [21], it is possible to observe general-purpose schemes, which usually includes entities of the type 'person', 'location', 'organization' and 'time', and schemes defined for specific applications. OntoNotes [22] is an example of the first type of approach: a broad collection of documents gathered from different sources (e.g., newspaper, television news) annotated with a tagset that includes general categories of named entities. On the other hand, more specific applications include biomedical NER, which focuses on identifying entities relevant to the biomedical field, such as diseases, genes and chemicals. An example in this field is the JNLPBA dataset[23], which is derived from the GENIA corpus. This dataset consists of 2,000 biomedical abstracts from the MEDLINE database, annotated with detailed entity types such as proteins, DNA, RNA, cell lines and cell types.

NER-based approaches for HS detection and analysis are still few. ElSherief et al. [24] exploited Twitter users' mentions to distinguish between directed and generalized forms of HS. Rodríguez-Sánchez et al. [25] used derogatory expressions of women as seeds to collect misogynist messages according to a fine grained classification of this phenomenon. [26] adopted a similar methodology to collect tweets about 3 vulnerable groups to discrimination: ethnic minorities, religious minorities, and Roma communities. Piot et al. [14] analyzed the correlation between the presence of HS and named entities in 60 existing datasets. Despite these previous works, there are no attempts to define a NER-based scheme specifically intended for HS detection. Our work represents an attempt to fill this gap by combining categories from general-purpose NER and a taxonomy of vulnerable groups to discrimination in a common annotation scheme aimed at providing deeper insights about the targets of HS.

3. Methodology

3.1. Data Collection

We collect news from public Telegram channels with the *telegram-dataset-builder* [27]. The selected channels are shown in Table 1, they are in Spanish and Italian and aligned with the left and right wings of the political spectrum. The subset of Italian headlines was integrated with titles published on newspapers Facebook pages that have been collected in collaboration with the Italian Amnesty Task Force on HS, a group of activists that produce counter narratives against discriminatory contents spread by online newspapers and users comments[2]. We collected all the news headlines detected by activists in March 2020, 2021, 2022, and 2023, and added them to our corpus.

Given the large amount of news collected, we applied filters to the dataset to reduce it to its final size. We focus on news about racism; for this purpose, we applied the classifier *piuba-bigdata/beto-contextualized-hate-speech* to stick to news items labeled as racism. Since this classifier is trained on Spanish

[2] https://www.amnesty.it/entra-in-azione/task-force-attivismo/

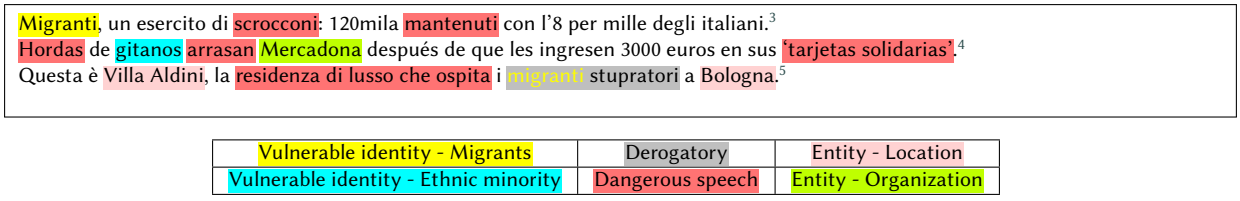

Figure 1: Examples of annotated headlines

	Left-wing	**Right-wing**
Spanish	elpais_esp, smolny7	MediterraneoDGT, elmundoes
Italian	ByobluOfficial, sadefenza	terzaroma, marcellopamio, ilprimatonazionaleIPN, VoxNewsInfo

Table 1
Telegram channels from which the news have been extracted.

texts, prior to this step we automatically translated Italian news with the model *facebook/nllb-200-distilled-600M*. This translation step is used only for the filtering process; once the news is selected, the translated text is no longer used. In the end, this process generates 532 news headlines classified as racist for Italian and 348 for Spanish, that have been selected for the annotation task.

3.2. Data Annotation

A comprehensive, span-based annotation scheme was developed to label vulnerable identities and entities present in the dataset. Annotators were provided with instructions and had to choose a label and highlight the word, phrase, or portion of text that best embodied the qualities of the chosen label in the text. It was possible to choose more than one label for the same portion of text. The instructions also provided annotators with some examples of annotated headlines.

The initial layer of annotation focuses on identifying vulnerable targets within the text and categorizing them into one of six predefined labels: **ethnic minority**, **migrant**, **religious minority**, **women**, **LGBTQ+ community**, and **other**. These labels represent vulnerable groups, as the vulnerability of the targets can often be traced back to their belonging to certain categories of people which are particularly exposed to discrimination, marginalisation, or prejudice in society. In cases where the targeted group didn't fit into one of the predefined labels, annotators were required to use the 'other' category. Then, for instances labeled as 'other', annotators were instructed to provide specific details regarding the group in a free-text field.

After categorizing vulnerable targets, the second layer involves annotating named entities. Annotators identify entities within the text and label them with one of five possible types: **person**, **group**, **organization**, **location**, and **other**. As in the first layer, instances labelled 'other' require annotators to provide details about the entity in a free-text field.

The final layers of the annotation scheme address the context in which these entities are mentioned, specifically focusing on identifying derogatory mentions and dangerous speech.

A derogatory mention is characterized by negative or disparaging remarks about the target. In these instances, explicit hate speech is absent, but the mention itself is discriminatory or offensive, often employing a tone intended to belittle or discredit the target. The label **derogatory** is used to mark these mentions.

Moreover, the annotation includes identifying dangerous elements: portions of text that, intentionally or unintentionally, could incite hate speech or increase the vulnerability of the target identity. Dangerous speech, which can be either explicit or implicit, promotes or perpetuates negative prejudices and stereotypes, potentially triggering harmful responses against the group. The label **dangerous** [28] is used to tag these segments. Annotators were encouraged to use free-text fields to provide details on implicit dangerous speech or recurring dangerous concepts.

The annotation guidelines provided annotators with specific criteria and with the following list of potential markers of dangerous speech to help their identification:

- **Incitement to violence**: the text explicitly encourages violence against the target group;
- **Open discrimination**: the text openly states or supports discrimination against the target group;
- **Ridicule**: the text ridicules the target in the eyes of the readers by belittling it or mocking it;
- **Stereotyping**: the text perpetuates negative stereotypes about the target group, contributing to a distorted view of it;
- **Disinformation**: the text spreads false or misleading information that can harm the target group;
- **Dehumanization**: the text dehumanizes the target group, using language that equates it with objects or animals;
- **Criminalization**: the text portrays the target group as inherently criminal or associates it with illegal activities, contributing to the perception that the group as a whole is dangerous.

However, a text may still be considered dangerous even if it does not explicitly include these markers, as they are intended as examples rather than strict requirements.

Figure 1 provides three examples of annotated headlines, two in Italian and one in Spanish, showing the application of the annotation scheme as described. In the figure, different colours highlight the various types of labels used. A vulnerable identity was detected in each headline: 'Migranti' in the first and in the third one and 'gitanos' in the second one, respectively labelled as 'vulnerable group - migrant' and

[2] "Migrants, an army of scroungers: 120,000 supported by the Italians' 8x1000 tax allocation".
[3] "Hordes of gypsies devastate Mercadona after 3000 euros were deposited in their solidarity cards".
[4] "This is Villa Aldini, the luxury residence that hosts rapist migrants in Bologna".

'vulnerable group - ethnic minority'. The three examples all contain multiple elements of dangerous speech, highlighted in red, and the second text also contains an element which was marked with the derogatory label. Additionally, the second and the third headlines include examples of annotation for named entities, with 'Mercadona' labelled as 'entity - organization', and 'Villa Aldini' and 'Bologna' labelled as 'entity - location'.

4. The VIRC Corpus

The VIRC corpus is a collection of 532 Italian and 348 Spanish news headlines annotated by 2 independent annotators for each language. Following the perspectivist paradigm [29], we both released the disaggregated annotations and the gold-standard corpus. The code used to generate the gold standard corpus, carry out experiments, and compile statistics can be accessed through the following GitHub repository[6]. In this Section we present an analysis of disagreement (Section 4.1) and relevant statistics about the corpus (Section 4.2).

4.1. Inter-Annotator Agreement

Since the span-based annotation task does not provide a fixed number of annotated items, we adopted the F-score metric to evaluate the agreement between annotators [30]. For each subset of the corpus we randomly chose one annotator as the gold standard set of labels and the other as the set of predictions. We then computed the F-score between the two distributions of labels in order to measure the agreement between the annotators. Table 2 shows the results of our analysis. In general, annotations always showed a fair or higher agreement, except for some entity-related labels and the "derogatory" one. There is also a low agreement in the Italian set on the labels "religious minority" and "women".

	IAA (F-score)	
	Spanish	**Italian**
dangerous	0.49	0.57
derogatory	0.08	0.28
entity - group	0.0	0.00
entity - location	0.66	0.60
entity - organization	0.41	0.12
entity - other	0.0	0.10
entity - person	0.47	0.63
vulnerable entity	0.15	0.00
vulnerable group - ethnic minority	0.83	0.63
vulnerable group - lgbtq+ community	-	0.80
vulnerable group - migrant	0.96	0.86
vulnerable group - other	0.46	0.41
vulnerable group - religious minority	1.0	0.00
vulnerable group - women	0.6	0.22

Table 2
The annotators agreement measured through the F-score and broken down by label.

Although the overall results are positive, they show significant variations that can be quantitatively and qualitatively. Inclusion of overlapping spans was handled as follows: if one span fully included another, this was considered to be an agreement. In cases where the spans only partially overlapped, meaning there was some shared text but not full inclusion, this was treated as a partial agreement. For example, if one annotator labeled "All women" and another selected only "women", this would be a full agreement (1 *true positive*). However, if the latter selected "women of Italy", it would be a partial agreement (0.5 *true positive*).

Quantitative Analysis. The agreement on the annotation of entities is always moderate but differs between the Spanish and the Italian subsets. Annotators of Spanish headlines scored a higher agreement on 'location' (0.66 *vs* 0.60), 'vulnerable' (0.15 *vs* 0) and 'organization' (0.41 *vs* 0.12) while entities of the type 'person' (0.63 *vs* 0.47) and 'other' (0.1 *vs* 0) are better recognized in Italian headlines.

On average, the annotation of vulnerable identities resulted in a higher agreement between annotators in both subsets and at the same time confirmed an higher agreement of Spanish annotations that always outperforms Italian ones. The highest agreement emerges for the label 'migrant' on which annotators obtained an F-score of 0.86 for Italian and 0.96 for Spanish. The agreement on 'ethnic minority' is a bit lower but still significant, while Spanish headlines reached an F-score of 0.83 Italian ones only 0.63. An equally high agreement is on the 'lgbtq+' label, which is only present in Italian headlines with an F-score of 0.8. Among vulnerable groups, women scored the lowest F-score: 0.6 for Spanish, 0.22 for Italian. The largest observed discrepancy is with religious minorities, in Spanish an F-score of 1 is achieved while in Italian 0.

While the annotation of 'dangerous' spans achieves an acceptable agreement, the 'derogatory' annotation is characterized as the one that achieves the lowest agreement between annotators. Additionally, annotations of Italian headlines resulted in higher disagreement than Spanish ones, contrary to what we observed about 'entities' and 'vulnerable identities'. Text spans expressing dangerous speech are recognized with an agreement of 0.57 for Italian and 0.49 for Spanish headlines. Agreement about 'derogatory' is low for Italian headlines (0.28) while Spanish ones show almost no agreement (0.08)

Qualitative Analysis. In summary, while the overall results of the annotation are positive, some categories show significant disagreement between annotators. These disagreements highlight the need to review and refine the annotation guidelines for problematic categories, and to provide more detailed instructions. The importance of reassessing the guidelines in order to make them clearer and more consistent is further underscored by the fact that, for Spanish headlines, the annotators agreed on both labels and intervals in only 67 cases, and for Italian headlines, agreement was reached in just 88 cases.

Since the annotation task was span-based, we opted not to use a confusion matrix to analyze the disagreement. A confusion matrix is not appropriate for span detection, as it assumes discrete labels applied to predefined items, whereas our task involved labeling spans of text that varied in length and context. Instead, we performed a qualitative analysis, examining specific cases of disagreement to understand their nature. This approach allowed us to explore not only how annotators differed in labeling spans but also why these differences emerged, providing a deeper insight into the underlying issues of interpretation and guidelines.

Looking more closely at the headlines where the annotations present inconsistencies, a variety of motivations behind discrepancies can be identified.

For instance, in the Italian title "Orrore nella casa occu-

[6] https://github.com/oeg-upm/virc

pata dagli immigrati: donna lanciata giù dal secondo piano"[7], 'donna' was marked as a vulnerable identity by only one of the annotators, suggesting maybe an erroneous focus on an individual target at a time ('immigrati') by the other annotator.

Another type of disagreement relates to the interpretation of derogatory mentions. An example can be found in "Un terzo dei reati sono commessi da stranieri (e gli africani hanno il record). Tutti i numeri"[8], where one annotator identified the term 'stranieri' as a derogatory mention, as well as representative of a vulnerable identity, while another annotator simply stuck to the second label, perhaps highlighting a divergence in the interpretation of the guidelines. Furthermore, it is interesting to observe the disagreement created by the headlines that use generic term 'stranieri' ('foreigners'), which was often labelled as 'vulnerable identity - ethnic minority' by one annotator and as 'vulnerable identity - migrant' by the other. This inconsistence between annotators can be identified in two headlines: "Ius soli e cittadinanza facile agli stranieri? Il sangue non è acqua"[9] and "Un terzo dei reati sono commessi da stranieri (e gli africani hanno il record). Tutti i numeri"[2]. In the first case, we can solve the disagreement by looking at the context: the explicit reference to the issue of granting citizenship suggests that the term 'foreigners' is more appropriately referred to the specific category of migrants. On the other hand, in the second headline, there is no direct reference to specifically migration-related issues and thus both interpretations in terms of the vulnerable category of belonging are acceptable.

Finally, some texts present a slight difference in the annotation spans of choice, as observed in "Più di 200mila case popolari agli immigrati"[10], where the annotators identified dangerous speech in the same section of text, but with differences in the number of highlighted words (first annotator labelled 'Più di 200mila'; second annotator labelled '200mila case popolari'), reflecting variations in the identification of relevant content for the analysis of dangerous speech.

In addition to the predefined labels, we also collected free-text fields as part of the annotation process. These comments offered an additional layer of granularity, allowing annotators to describe nuances not covered by the fixed categories. For example, in the Spanish headline "Dos menas marroquíes apuñalan a dos turistas para robarles en Salou"[11], both annotators used the two labels 'vulnerable identity - ethnic minority' and 'vulnerable identity - other' to annotate the span 'menas marroquíes'. Alongside the 'other' label, one annotator provided the comment 'Under 18', while the other one used 'young people' to describe the vulnerable group. Although stated differently, both comments highlight the specific vulnerability related to the age of the group, complementing the existing labels. As this example shows, the flexibility in the annotation process provided by free-text fields is useful to capture multi-categorical terms and to identify potential new categories that may not have been initially considered in the predefined labels.

[7] "Atrocity in a house occupied by migrants: woman thrown from second floor".
[8] "One third of all crimes are committed by foreigners (and Africans hold the record). All the numbers".
[9] "Ius soli and easy citizenships for foreigners? Blood is not water".
[10] "More than 200,000 public housing units for immigrants".
[11] "Two Moroccan unaccompanied migrant minors stab two tourists to rob them in Salou".

	Spanish	Italian
dangerous	136	166
derogatory	3	16
entities	140	146
vulnerable groups	270	253

Table 3
The distribution of labels in the gold standard corpus.

4.2. Dataset Analysis

In this section we provide an analysis of the four label types that occur in the gold standard version of the VIRC corpus: 'derogatory', 'dangerous', 'named entities', 'vulnerable groups'. The analysis is twofold: first, we describe the distribution of these label types, then we present a zero-shot and a few-shot experiment aimed at understanding if existing LLMs (T5[31] and BART[32]) are able to recognize these labeled spans in news headlines by comparing their outputs to the gold standard annotations.

Corpus statistics. Table 3 shows the distribution of label types in the corpus. As it can be observed, mentions of vulnerable groups are the most present, with 270 occurrences in the Spanish subset and 253 in the Italian subset. This confirms the relevance of annotating vulnerable in the identification of discriminatory contents, which is tied to their high recognizability by annotators (Section 4.1). The role on named entities differs in the two subsets. Annotators labeled them with agreement 130 times in Spanish headlines and 67 times in Italian ones. This might be caused by their compositions. Since Italian headlines were partly collected from Facebook pages of mainstream newspapers, there was a higher number of named entities that were not relevant for the analysis of headlines' danger. The number of text spans labeled as dangerous is almost equivalent in the two subsets (136 for Spanish, 166 for Italian), showing a good presence of this label type despite the high disagreement between annotators. Finally, it is worth mentioning the almost total absence of text spans labeled as 'derogatory' with agreement (3 for Spanish, 16 for Italian) that suggests the high subjectivity of this phenomenon and also the need of better define its characteristics in annotation guidelines.

Corpus analysis with LLMs. We completed our analysis of the VIRC corpus through zero-shot experiments aimed at exploring the ability of existing LLMs to identify the four types of labelled spans in messages. We considered the detection of spans as an extractive Question Answering (QA) problem. For the task we adopted the T5[31] and BART[32] LLMs architectures for both languages. For Italian we employ [33] and [34] and for Spanish [35] and [36] models, respectively. The translations of the prompts used are the following (see Appendix A for the original ones):

- What part of the text is dangerous (criminalizes, ridicules, incites violence, ...) against vulnerable identities (women, migrants, ethnic minorities, ...)?
- What part of the text is derogatory (negative or pejorative comments about the victim without explicit hate speech, but the mention itself is discriminatory or offensive, and often uses a tone intended to denigrate or discredit the victim)?
- What named entity is mentioned in the sentence?

	Non-Restictive Zero-Shot				Restictive Zero-Shot			
	T5		BART		T5		BART	
	Spanish	Italian	Spanish	Italian	Spanish	Italian	Spanish	Italian
dangerous	0.39	0.28	0.43	0.39	0.49	0.47	0.51	0.43
derogatory	0.02	0.05	0.03	0.04	0.67	0.43	0.50	0.33
entity	0.28	0.11	0.23	0.23	0.40	0.30	0.30	0.27
vulnerable identity	0.63	0.19	0.41	0.48	0.56	0.18	0.35	0.37

Table 4
F-score results of zero-shot experiments on the VIRC corpus with T5 and BART models for each label.

- Which hate speech vulnerable identity is mentioned in the sentence?

We designed two approaches for zero-shot experiments, restictive and non-restrictive. On the one hand, for the **non-restictive zero-shot** experiments, for each sentence in the dataset, we queried the model with the prompt of each label and extracted the three most confident results. Then, we filtered out those responses below the $\%0.02$ confidence of the model to limit the noise. Finally, all these annotations go through a majority vote (identical to the one used to build the aggregate dataset) to normalize the model response.

On the other hand, for the **restictive zero-shot** experiments, we queried the model with the prompts for each annotation present in the aggregated dataset. And, as there are sentences that have two equal labels in different spans, we request five different annotations from the model, ordered from most confident to least confident. If an annotation was already included, the next annotation is taken in order to avoid duplicating annotations in the model.

Table 4 presents the F-scores for each label type, experiment, and model. In general, T5 and BART tend to perform more effectively in Spanish compared to Italian. The models face noticeable challenges in identifying the labels 'dangerous', 'derogatory', and 'entity'. Nevertheless, when they are aware that the label exists within the sentence (restictive), they manage to recognize it with fairly good agreement. During annotation, the label 'derogatory' proves most challenging to identify. In the non-restrictive scenario, it scarcely receives any agreement, yet in the restictive scenario, it achieves a reasonable level, particularly in Spanish. This indicates that the model struggles to discern its presence initially but, once acknowledged, can recognise the expression.

The restictive method enhances performance over the non-restictive method for all labels except 'vulnerable identity.' This shows that models generally have a better comprehension and identification of vulnerable identities in sentences without restrictions compared to when they are restricted to specific mentions. It should also be noted that, in the Spanish context, T5 is more effective than BART in identifying 'vulnerable identity' labels for both approaches, while BART performs better in Italian.

These results show that a NER-based annotation scheme for HS detection is difficult to annotated but also to be automatically detected. Larger resources are necessary to develop models that are able to detect the complex semantics of HS.

5. Conclusions and Future Work

The Vulnerable Identities Recognition Corpus (VIRC), created in this work, reveals the challenge of identifying vulnerable identities due to the rapid evolution of language on social media. Our experiments indicate that large language models (LLMs) struggle significantly with this task.

VIRC provides a detailed and structured resource that enhances understanding of the extensive use of hate speech in Italian and Spanish news headlines. The corpus is particularly valuable as it includes more annotation dimensions compared to related studies in other languages, such as vulnerable identities, dangerous discourse, derogatory expressions, and entities. This differentiation between vulnerable identities and entities, as well as between dangerous and derogatory elements, enables the development of sophisticated detection tools that can facilitate large-scale actions to mitigate the impact of hate speech (e.g., moderation of messages and generation of counter-narratives that reduce the damage to the mental health of victims).

Future work will focus on expanding this resource by doubling the size of annotations for both languages and including non-racism-related phrases to ensure the resource is comprehensive. Additionally, we plan to refine the annotation guidelines to avoid low agreement on the derogatory label, enhancing the overall reliability and utility of the corpus. These efforts will further improve the effectiveness of hate speech detection and contribute to the development of policies and tools for a safer online environment.

Acknowledgments

This work is supported by the Predoctoral Grant (PIPF-2022/COM-25947) of the Consejería de Educación, Ciencia y Universidades de la Comunidad de Madrid, Spain. Arianna Longo's work has been supported by aequa-tech. The authors gratefully acknowledge the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on the IPTC-AI innovation Space AI Supercomputing Cluster.

References

[1] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 512–515.

[2] Z. Waseem, Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter, in: Proceedings of the first workshop on NLP and computational social science, 2016, pp. 138–142.

[3] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, D. Yang, Latent hatred: A benchmark for understanding implicit hate speech, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 345–363.

[4] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, R. Tromble, Introducing cad: the contextual abuse dataset, in: Proceedings of the 2021 Conference of the

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2289–2303.

[5] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, V. Patti, Emotionally informed hate speech detection: A multi-target perspective, Cogn. Comput. 14 (2022) 322–352. URL: https://doi.org/10.1007/s12559-021-09862-5. doi:10.1007/S12559-021-09862-5.

[6] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1668–1678.

[7] P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. Von Vacano, C. Kennedy, The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022, 2022, pp. 83–94.

[8] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 14867–14875.

[9] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, Semeval-2021 task 5: Toxic spans detection, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), 2021, pp. 59–69.

[10] K. Büyükdemirci, I. E. Kucukkaya, E. Ölmez, C. Toraman, JL-Hate: An Annotated Dataset for Joint Learning of Hate Speech and Target Detection, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9543–9553.

[11] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Lang. Resour. Evaluation 55 (2021) 477–523. URL: https://doi.org/10.1007/s10579-020-09502-8. doi:10.1007/S10579-020-09502-8.

[12] E. Leonardelli, S. Menini, A. P. Aprosio, M. Guerini, S. Tonelli, Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 10528–10539.

[13] H. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2193–2210.

[14] P. Piot, P. Martín-Rodilla, J. Parapar, Metahate: A dataset for unifying efforts on hate speech detection, Proceedings of the International AAAI Conference on Web and Social Media 18 (2024) 2025–2039. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/31445. doi:10.1609/icwsm.v18i1.31445.

[15] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate speech detection in Italian social media text, in: K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, Z. Talat (Eds.), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 252–260. doi:10.18653/v1/2022.woah-1.24.

[16] M. Madeddu, S. Frenda, M. Lai, V. Patti, V. Basile, Disaggreghate it corpus: A disaggregated italian dataset of hate speech, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)., volume 3596, 2023.

[17] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, SemEval-2021 task 5: Toxic spans detection, in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 59–69. URL: https://aclanthology.org/2021.semeval-1.6. doi:10.18653/v1/2021.semeval-1.6.

[18] P. G. Hoang, C. D. Luu, K. Q. Tran, K. V. Nguyen, N. L.-T. Nguyen, ViHOS: Hate speech spans detection for Vietnamese, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 652–669. URL: https://aclanthology.org/2023.eacl-main.47. doi:10.18653/v1/2023.eacl-main.47.

[19] Y. Jeong, J. Oh, J. Lee, J. Ahn, J. Moon, S. Park, A. Oh, KOLD: Korean offensive language dataset, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10818–10833. URL: https://aclanthology.org/2022.emnlp-main.744. doi:10.18653/v1/2022.emnlp-main.744.

[20] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, D.-Y. Yeung, Multilingual and multi-aspect hate speech analysis, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4675–4684. URL: https://aclanthology.org/D19-1474. doi:10.18653/v1/D19-1474.

[21] B. Jehangir, S. Radhakrishnan, R. Agarwal, A survey on named entity recognition - datasets, tools, and methodologies, Natural Language Processing Journal 3 (2023).

[22] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, R. Weischedel, Ontonotes: the 90% solution, in: Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers, 2006, pp. 57–60.

[23] N. Collier, T. Ohta, Y. Tsuruoka, Y. Tateisi, J.-D. Kim, Introduction to the bio-entity recognition task at jnlpba, in: N. Collier, P. Ruch, A. Nazarenko (Eds.), Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), COLING, 2004, pp. 73–78.

[24] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, E. Belding, Hate lingo: A target-based linguistic analysis of hate speech in social media, in: Proceedings of the international AAAI conference on web and social media, volume 12, 2018.

[25] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576.

[26] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An italian twitter corpus of hate speech against immigrants, in: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 2018.
[27] I. Guillén-Pacho, oeg-upm/telegram-dataset-builder: version 1.0.0, 2024. URL: https://doi.org/10.5281/zenodo.12773159. doi:10.5281/zenodo.12773159.
[28] S. Benesch, Dangerous speech, 86272 12 (2023) 185–197.
[29] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 6860–6868.
[30] T. Brants, Inter-annotator agreement for a german newspaper corpus., in: LREC, Citeseer, 2000.
[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.
[32] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL: https://arxiv.org/abs/1910.13461. arXiv:1910.13461.
[33] G. Sarti, M. Nissim, IT5: Text-to-text pretraining for Italian language understanding and generation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9422–9433. URL: https://aclanthology.org/2024.lrec-main.823.
[34] M. La Quatra, L. Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2023). URL: https://www.mdpi.com/1999-5903/15/1/15. doi:10.3390/fi15010015.
[35] V. Araujo, M. M. Trusca, R. Tufiño, M.-F. Moens, Sequence-to-sequence spanish pre-trained language models, 2023. arXiv:2309.11259.
[36] V. Araujo, M. M. Trusca, R. Tufiño, M.-F. Moens, Sequence-to-sequence spanish pre-trained language models, 2023. arXiv:2309.11259.

A. LLMs Prompts

The prompts used are the same for each model but different for each language. For Spanish, the prompts used for each label are:

- **Dangerous:** "¿Qué parte del texto es peligroso (criminaliza, ridiculiza, incita a la violencia, ...) contra identidades vulnerables (mujeres, migrantes, minorías étnicas, ...)?"
- **Derogatory:** "¿Qué parte del texto es derogativo (comentarios negativos o despectivos sobre la víctima sin incitación explícita al odio, pero la mención en sí es discriminatoria u ofensiva, y a menudo emplea un tono destinado a menospreciar o desacreditar a la víctima)?"
- **Entity:** "¿Qué entidad nombrada se menciona en la frase?"
- **Vulnerable Identity:** "¿Qué identidad vulnerable al discurso de odio se menciona en la frase?"

For Italian:

- **Dangerous:** "Quale parte del testo è pericolosa (criminalizza, ridicolizza, incita alla violenza, ...) nei confronti di identità vulnerabili (donne, migranti, minoranze etniche, ...)?"
- **Derogatory:** "Quale parte del testo è dispregiativa (commenti negativi o denigratori sulla vittima senza un esplicito discorso d'odio, ma in cui la menzione stessa è discriminatoria o offensiva e spesso usa un tono volto a sminuire o screditare la vittima)?"
- **Entity:** "Quale entità nominata è menzionata nella frase?"
- **Vulnerable Identity:** "Quale identità vulnerabile ai discorsi d'odio è menzionata nella frase?"