The Vulnerable Identity Recognition Corpus Annotation Guidelines

Introduction

Vulnerable Identity Recognition is a task aimed at recognising vulnerable identities in journalistic writing. The task is based on two subtasks:

1. The first one consists in identifying **vulnerable identities** and **named entities**;
2. The second one is to recognise whether the text in which they are mentioned has a **discriminatory** nature or not and whether it contains elements that are **dangerous** in relation to them.

Vulnerable targets are groups or individuals that may be particularly susceptible to prejudice, discrimination, or harm, especially if their identity is exposed or treated insensitively. These may include ethnic and religious minorities, LGBTQ+ people, people with disabilities, victims of violence or abuse, minors, young people, people with mental health problems and others.

The dataset consists of 543 Italian and 378 Spanish news headlines collected from Facebook posts and Telegram groups. The aim of the resource is to enhance hate speech analysis in Italian and Spanish news headlines.

Annotation scheme

The annotation scheme of the task consists of several layers, and it is span-based. In order to annotate, annotators must choose a label and highlight the word, phrase, or section of text that best embodies the qualities of the chosen label in the text. It is possible to choose more than one label for the same portion of text.

Vulnerable identities

The first layer of the annotation scheme refers to vulnerable identities and to vulnerable categories, as the vulnerability of the targets can often be traced back to their belonging to certain groups of people which are particularly exposed to discrimination, marginalisation or prejudice in society.

Six labels are provided to indicate the vulnerable category the target belongs to:

- **Ethnic minority**
- **Migrants**
- **Religious minority**
- **Women**
- **LGBTQ+ community**
- **Other**

For instances labelled as ‘other’, annotators are required to provide specific details regarding the entity in a free-text field.

Entities

The second layer involves annotating named entities. Five labels are provided to describe the type of entity mentioned in the title:

- **Person**: a single person or identifiable individual.
- **Group**: any set of individuals who share a common characteristic or are brought together for a specific purpose.
- **Organization**: any formal or structured entity that operates with a specific purpose, such as companies, institutions, non-profit organisations, or governmental bodies.
- **Location**: specific geographical areas; countries, cities, buildings, or any other place.
- **Other**

For instances labelled as 'other', annotators are required to provide specific details regarding the entity in a free-text field.

Derogatory mention

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.

Annotators use the label **derogatory** to mark these mentions.

Dangerous speech

Any form of expression that could, intentionally or unintentionally, **incite hate speech** or increase the vulnerability of the target identity. **Dangerous speech**, which can be either explicit or implicit, promotes or perpetuates prejudice and negative stereotypes, potentially triggering harmful responses against the group.

Annotators apply the label **dangerous** to these segments and can use free-text fields to provide details on implicit dangerous speech or recurring dangerous concepts.

A headline can be considered dangerous when it contains one or more of the following markers:

- **Incitement to violence**: the text explicitly encourages violence against the target group;
- **Open discrimination**: the text openly state or support 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 them 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 seen as dangerous even if it does not explicitly include these markers, as they are intended as examples rather than strict requirements.

Examples of annotated headlines

"Coppie gay spacciano i bimbi per figli"[1]

- coppie gay → vulnerable identity: LGBTQ+ community
- spacciano i bimbi → dangerous

"Migranti, un esercito di scrocconi: 120,000 mantenuti con l'8xmille degli italiani"[2]

- Migranti → vulnerable identity: migrants
- un esercito di scrocconi → dangerous
- mantenuti → dangerous

"Detenido un cubano rabioso por morder a su mujer: '¡Eres una puta!'"[3]

- cubano → vulnerable identity: ethnic minority
- cubano rabioso → derogatory
- Detenido → dangerous
- por morder → dangerous

[1] "Gay couples pretend children are their own".
[2] "Migrants, an army of scroungers: 120,000 supported by the Italians' 8x1000 tax allocation".
[3] "Rabid Cuban arrested for biting his wife: 'You're a whore!'".