Annotation guidelines for AmnestyCounterHS

Detection task of dangerous speech and counter-narratives

Introduction

The focus of this research is **counter-narratives**. The aim of this task is to recognise what strategies of counter-narratives are the most used in Italy, also by referring to the target addressed by the hateful comment.

The dataset consists of 307 comments written by Facebook users and Amnesty International activists, posted under 20 public Facebook posts that featured some news titles. It includes both responses to other users' comments and direct replies to the posts themselves. In the annotation outline, the annotator can see both the title and the comment posted beneath it.

The annotation scheme was developed based on two pre-existing works:

- CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech
- Counter-TWIT: An Italian Corpus for Online Counterspeech in Ecological Contexts

This annotation scheme integrates and maps the annotation schemes used in these two cited papers.

The annotation is made up of four layers:

1. In the first layer the annotator has to specify if the text you are reading is written using a formal or informal **style**.
2. In the second layer the annotator has to identify if the comment is **supporting** another dangerous speech or counterspeech comment.
3. In the third layer the annotator has to identify if there is **dangerous speech** in the comment and specify if it is explicit or implicit.
4. In the fourth layer the annotator has to identify if the comment is a **counter-narrative** and which counter-narrative **strategy** has been used.

1. What's the textual style of the comment?

- **Formal** style → a formal text contains the following characteristics:
  - Third person;
  - More complex vocabulary.
- **Informal** style → an informal text contains the following characteristics:
  - First person;
  - Simpler words;
  - Abbreviations;
  - Slang;
  - Bad words.

2. Can you see any support in the comment?

In this layer the annotator is asked to highlight the portion of the text where he/she can see any support. The support can be of two types:

- Dangerous speech support:

  > “a parte gli scherzi (sono d'accordo con lei) sappiamo bene che se arriva qualche barcone li accogliamo come sempre fatto, non ci nascondiamo dietro un dito”

- Counterspeech support:

  > “Come dice giustamente il signor Mario, comportamenti di questo tipo sono frequenti in tutti i regimi dittatoriali, in cui i cittadini non possono decidere liberamente della propria esistenza.”

3. Can you see any dangerous speech in the comment? → YES or NO

Dangerous speech is any form of expression that **promotes** or perpetuates negative **prejudices** and **stereotypes against certain social groups** based on characteristics such as gender, ethnicity, age, sexual orientation, religion, or disability.

These texts can spread hate, exclusion, and marginalisation towards the targeted groups.

If you can see dangerous speech in the comment we ask you to identify the type of that dangerous speech and to highlight the portion of the text that made it clear for you.

- Explicit

  "Mandatelo a casa"

- Implicit

  News title: "Addio a Lucy Salani, era la donna transessuale più anziana d'Italia"

  Comment: "Maaa… esistono ancora persone normali?" → In this case the person who wrote the comment indirectly says that transexual people are not normal

4. Can you see any counter narrative in the comment? → YES or NO

A possible solution to fight online dangerous speech is the generation of counter-narratives. **Counter-narratives** are a form of counterspeech used to fight online hate speech by directly responding to those who post hateful comments and trying to get them to reconsider their actions, or try to convince other people that what has been said by the user is wrong.

If you can see any counterspeech in the comment we ask you to identify the **strategy** of counterspeech that has been used.

There are several strategies of counter-narratives that can be used to fight online hate speech:

- **Informative**: the writer writes a statement that seeks to debunk or fact-check the claims made by the perpetrator.

  "Minoranze etniche è un termine usato in un contesto specifico, qui ad esempio, nel Regno Unito, le persone provenienti da questi paesi sono minoranze".

- **Alternative**: the writer proposes alternatives to the statement made by the perpetrator and proposes corrections about some aspects of its content, suggesting a more “correct” point of view and giving a more detailed description of facts.

  "Non gigante buono, ma femminicida"

- **Suggestion**: the writer suggests actions to the perpetrator to encourage them to rethink their views.

  "Le consiglio di leggere degli articoli sull’argomento"

- **Explicitation**: the writer explicitates/reveals what was implicit in the statement made by the perpetrator.

  “Stanno equiparando la pedofilia all'omosessualità”

- **Question**: questions that would challenge the speaker’s chain of reasoning and compel them to either answer convincingly or recant their original remark.

  “Si potrebbe almeno riportare qualche fatto prima di trarre queste conclusioni?”

  Indirect questions should be annotated too.

  “mi dia qualche link che riporti esempi concreti di quanto afferma”

- **Denouncing and explaining**: when you convey the impression that the opinions put forth by the hate speaker are not acceptable and you try to explain to the user why.

  “C'è un grosso errore di fondo in quanto scritto nell'introduzione di questo articolo. Rendere l'interruzione di gravidanza un diritto garantito dall'assistenza sanitaria pubblica non significa che lo Stato imponga alcunché.”

- **Positive**: a courteous, polite, and civil statement.

  “Insegnare ai bambini che ci sono tanti modi differenti per essere felici e che i loro sentimenti valgono è una cosa su cui concordo totalmente.”

- **Hostile**: the user expresses hostility, aggressiveness towards the initial content, using insults or aggressive words.

  “Bisogna davvero essere degli stupidi idioti retrogradi a credere alla negatività sull’Islam.”

- **Humour**: a strategy of counterspeech with an humoristic, ironic, sarcastic intent whether positive or negative.

"E meno male che era buono. Se era cattivo che faceva, se la magnava?"

It is possible to identify more than a single counterspeech strategy in a single comment.

References

Dangerous Speech: A Practical Guide by the Dangerous Speech Project

Counter-TWIT: An Italian Corpus for Online Counterspeech in Ecological Contexts, Pierpaolo Goffredo, Valerio Basile, Bianca Cepollaro, Viviana Patti

CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech, Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini, 2019.

Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation, Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhkavi, Tanmoy Chakraborty, and Md Shad Akhtar