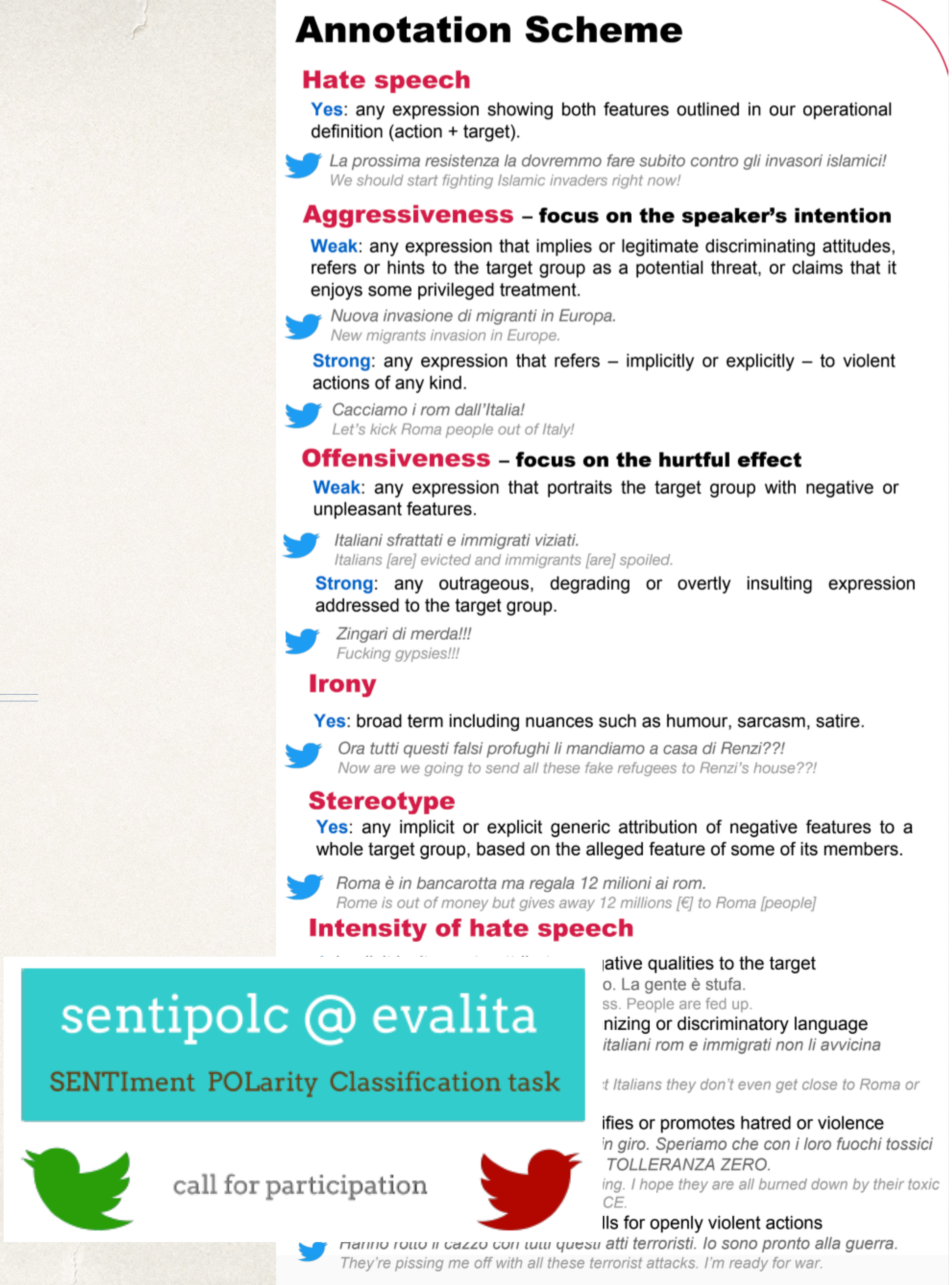
Developing corpora for sentiment analysis / hate speech detection, …



Annotation Scheme

Hate speech

Yes: any expression showing both features outlined in our operational definition (action + target).

La prossima resistenza la dovremmo fare subito contro gli invasori islamici!
We should start fighting Islamic invaders right now!

Aggressiveness – focus on the speaker's intention

Weak: any expression that implies or legitimate discriminating attitudes, refers or hints to the target group as a potential threat, or claims that it enjoys some privileged treatment.

Nuova invasione di migranti in Europa.
New migrants invasion in Europe.

Strong: any expression that refers – implicitly or explicitly – to violent actions of any kind.

Cacciamo i rom dall'Italia!
Let's kick Roma people out of Italy!

Offensiveness – focus on the hurtful effect

Weak: any expression that portraits the target group with negative or unpleasant features.

Italiani sfrattati e immigrati viziati.
Italians [are] evicted and immigrants [are] spoiled.

Strong: any outrageous, degrading or overtly insulting expression addressed to the target group.

Zingari di merda!!!
Fucking gypsies!!!

Irony

Yes: broad term including nuances such as humour, sarcasm, satire.

Ora tutti questi falsi profughi li mandiamo a casa di Renzi??!
Now are we going to send all these fake refugees to Renzi's house??!

Stereotype

Yes: any implicit or explicit generic attribution of negative features to a whole target group, based on the alleged feature of some of its members.

Roma è in bancarotta ma regala 12 milioni ai rom.
Rome is out of money but gives away 12 millions [€] to Roma [people]

Intensity of hate speech

ative qualities to the target
o. La gente è stufa.
ss. People are fed up.
nizing or discriminatory language
italiani rom e immigrati non li avvicina
t Italians they don't even get close to Roma or
ifies or promotes hatred or violence
in giro. Speriamo che con i loro fuochi tossici
TOLLERANZA ZERO.
ing. I hope they are all burned down by their toxic
CE.
lls for openly violent actions
Hanno rotto il cazzo con tutti questi atti terroristi. Io sono pronto alla guerra.
They're pissing me off with all these terrorist attacks. I'm ready for war.

sentipolc @ evalita

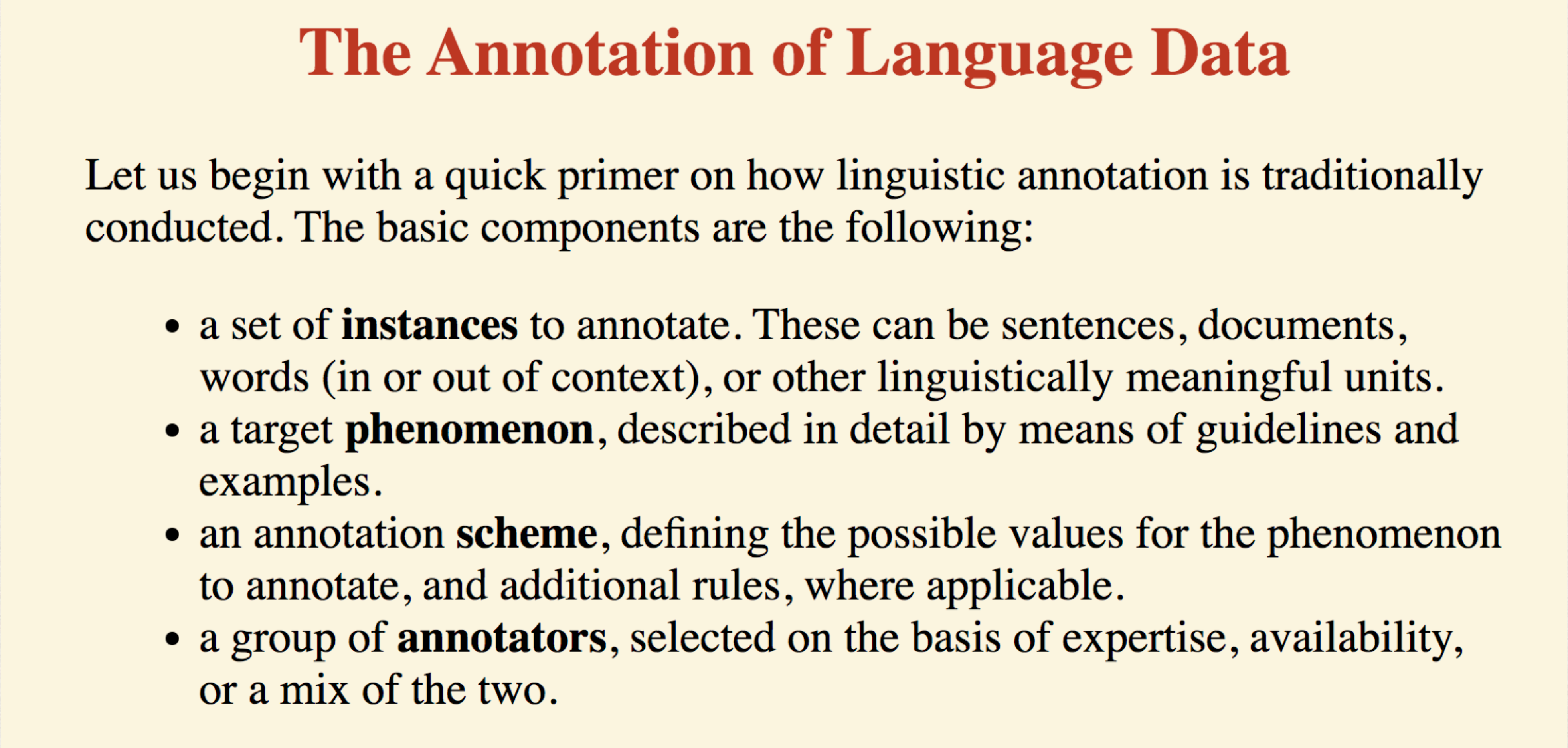
SENTIment POLarity Classification task

call for participation

Annotated corpora for SA et. Al

- The most used resources for NLP are currently annotated corpora, where linguistic data are associated with explicit annotation of the most relevant part of linguistic knowledge.
- Corpora have been developed during the last decades for a variety of NLP tasks:
    - corpora for sentiment analysis, where information concerning the polarity of linguistic expressions or sentences is made explicit

Annotated corpora for SA et al.



The Annotation of Language Data

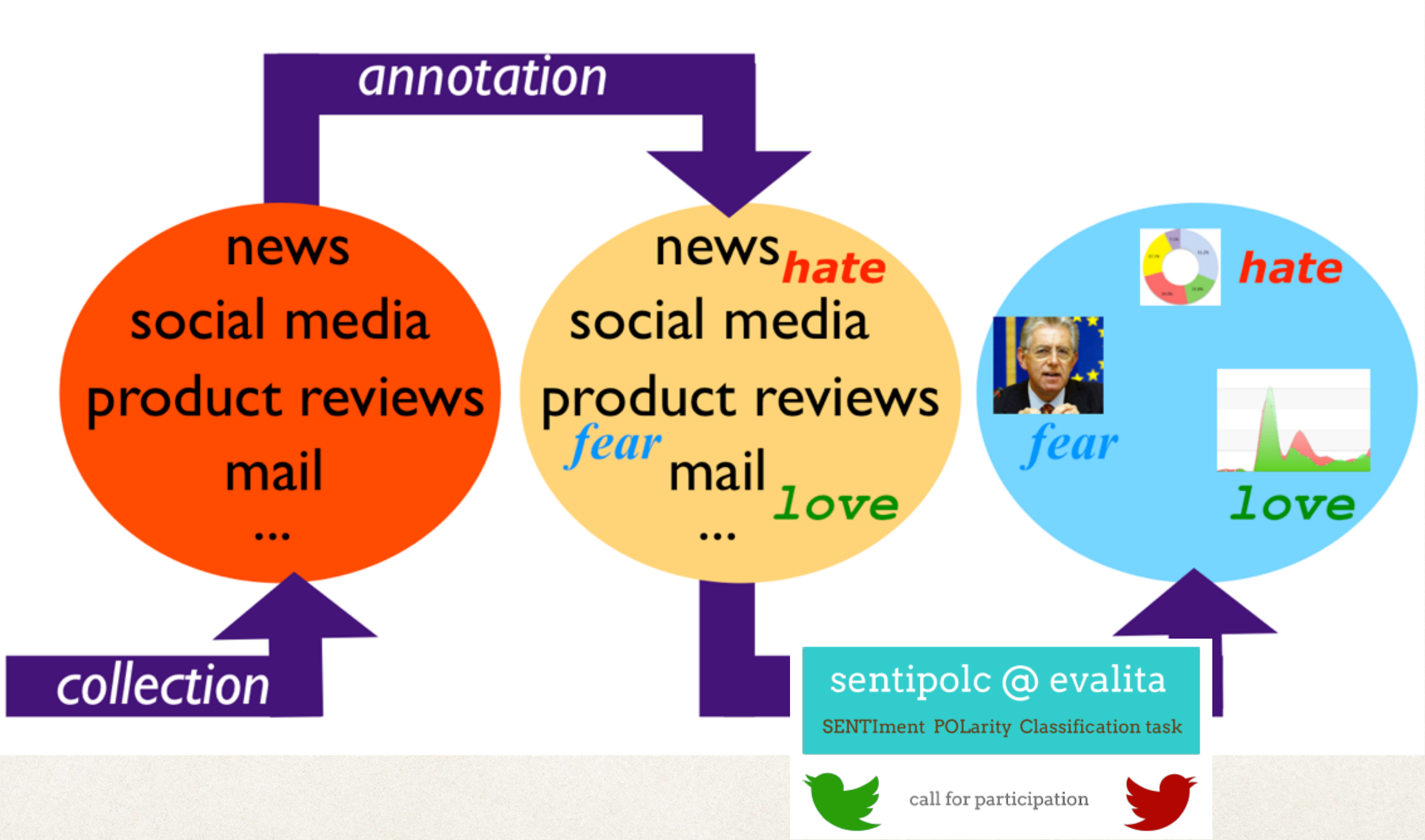
Let us begin with a quick primer on how linguistic annotation is traditionally conducted. The basic components are the following:

- a set of **instances** to annotate. These can be sentences, documents, words (in or out of context), or other linguistically meaningful units.
- a target **phenomenon**, described in detail by means of guidelines and examples.
- an annotation **scheme**, defining the possible values for the phenomenon to annotate, and additional rules, where applicable.
- a group of **annotators**, selected on the basis of expertise, availability, or a mix of the two.

Annotated corpora for SA et al.

- Annotated corpora, where linguistic data are associated with:
    - explicit annotation of the most relevant part of linguistic knowledge for the task of interest
- In general, the development of a linguistic resource includes:
    - collection of data to be annotated (data balance, copyright solution)
    - definition of an annotation scheme to be applied (what kind of information, what kind of representation and format)
    - application of the scheme to data (manually, involving a wide/diverse set of competent humans or automatically)
    - validation of the annotated data
        - Agreement/disagreement metrics, comparison, system training

Annotated corpora for SA et al.



annotation

news
social media
product reviews
mail
…

news
social media
product reviews
mail
…
hate
fear
love

hate
fear
love

collection

sentipolc @ evalita
SENTIment POLarity Classification task
call for participation

Exploitation in training / fine-tuning and testing automatic systems

Annotated corpora for SA et al.

- In corpora developed for sentiment analysis the collection usually
    - are focused on social media, blogs, site where posts comment about politics, products…
    - is done according to the policies stated by providers
    - includes data which can be considered as a statistically representative of the phenomena to be studied
        - the importance of a good sample!

Developing corpora for hate speech

- Selecting data samples
    - Collect data from sources representative of the phenomena to be studied
    - Filter data by keywords and hashtags representing:
        - Hate speech targets > e.g. women, immigrants (Romas, Muslims, …)
        - Forms of hate speech > misogyny, racism, xenophobia, religious hate..
        - Monitoring potential victims of hate accounts, downloading the history of identified haters and filtering Twitter streams with keywords, i.e. words, hashtags and stems.
        - Media ecosystem (reactions to news posts)

Annotated corpora for SA et al.

- In corpora developed for sentiment analysis the annotation scheme is oriented to made explicit
    - the polarity of each post (is the sentiment/opinion expressed positive or negative?, …)
        - or other labels depending on the focus of the task (sentiment polarity, emotions, stance, hate speech, …)
    - the entity towards which the sentiment/opinion is expressed (target)
    - the presence of figurative use of language (irony, metaphor, …)
    - …

Annotated corpora for SA et. al.

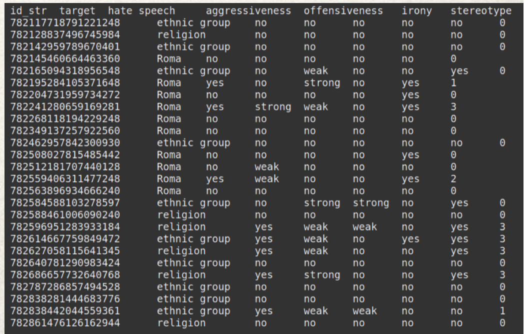
- Testing the accuracy of automatic systems in classifying the text according to a sentiment scheme requires the availability of a manually annotated dataset where the sentiment in the texts has been classified by several human experts
- Application of the annotation scheme:
  - manually or semi-automatically
    - manually: by at least 3 skilled human annotators
    - crowd vs experts
    - annotation guidelines

Crowdsourcing
Annotation platforms

- Amazon Mechanical Turks: https://www.mturk.com/
- Prolific: https://www.prolific.com/
- Appen (ex Crowdflower): https://www.appen.com/
- Label studio: https://labelstud.io/
- Home made platforms
- …

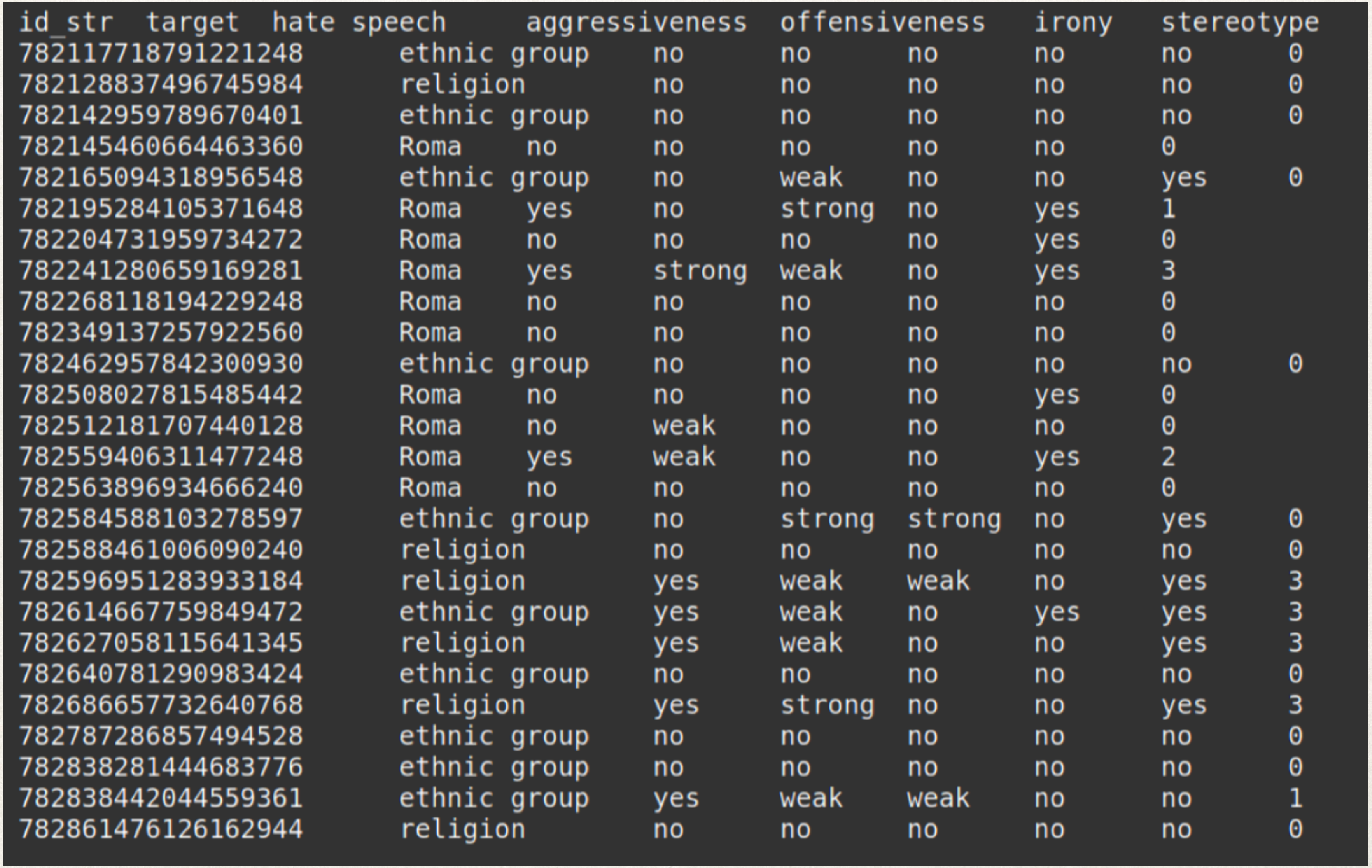
Developing corpora for hate speech

- Annotation scheme applied by human annotators/judges (expert vs crowdsourcing)
    - Labels oriented to made explicit the presence of hate speech in texts , given an operational definition)
            - Coarse-grained: Hateful? Yes or no;.Misogyny? Yes or no
            - Fine-grained: relevant aspects characterizing hate
        - The entity towards which the hate is expressed (target)
        - Presence of figurative use of language: irony/sarcasm
        - Multilayered annotation schemes



```
id_str  target  hate speech     aggressiveness  offensiveness   irony   stereotype
782117718791221248      ethnic group    no      no      no      no      no      0
782128837496745984      religion        no      no      no      no      no      0
782142959789670401      ethnic group    no      no      no      no      no      0
782145460664463360      Roma    no      no      no      no      no      0
782165094318956548      ethnic group    no      weak    no      no      yes     0
782195284105371648      Roma    yes     no      strong  no      yes     1
782204731959734272      Roma    no      no      no      no      yes     0
782241280659169281      Roma    yes     strong  weak    no      yes     3
782268118194229248      Roma    no      no      no      no      no      0
782349137257922560      Roma    no      no      no      no      no      0
782462957842300930      ethnic group    no      no      no      no      no      0
782508027815485442      Roma    no      no      no      no      yes     0
782512181707440128      Roma    no      weak    no      no      no      0
782559406311477248      Roma    yes     weak    no      no      yes     2
782563896934666240      Roma    no      no      no      no      no      0
782584588103278597      ethnic group    no      strong  strong  no      yes     0
782588461006090240      religion        no      no      no      no      no      0
782596951283933184      religion        yes     weak    weak    no      yes     3
782614667759849472      ethnic group    yes     weak    no      yes     yes     3
782627058115641345      religion        yes     weak    no      no      yes     3
782640781290983424      ethnic group    no      no      no      no      no      0
782686657732640768      religion        yes     strong  no      no      yes     3
782787286857494528      ethnic group    no      no      no      no      no      0
782838281444683776      ethnic group    no      no      no      no      no      0
782838442044559361      ethnic group    yes     weak    weak    no      no      1
782861476126162944      religion        no      no      no      no      no      0
```

Developing corpora for hate speech



```
id_str  target  hate speech     aggressiveness  offensiveness   irony   stereotype
782117718791221248      ethnic group    no      no      no      no      no      0
782128837496745984      religion        no      no      no      no      no      0
782142959789670401      ethnic group    no      no      no      no      no      0
782145460664463360      Roma    no      no      no      no      no      0
782165094318956548      ethnic group    no      weak    no      no      yes     0
782195284105371648      Roma    yes     no      strong  no      yes     1
782204731959734272      Roma    no      no      no      no      yes     0
782241280659169281      Roma    yes     strong  weak    no      yes     3
782268118194229248      Roma    no      no      no      no      no      0
782349137257922560      Roma    no      no      no      no      no      0
782462957842300930      ethnic group    no      no      no      no      no      0
782508027815485442      Roma    no      no      no      no      yes     0
782512181707440128      Roma    no      weak    no      no      no      0
782559406311477248      Roma    yes     weak    no      no      yes     2
782563896934666240      Roma    no      no      no      no      no      0
782584588103278597      ethnic group    no      strong  strong  no      yes     0
782588461006090240      religion        no      no      no      no      no      0
782596951283933184      religion        yes     weak    weak    no      yes     3
782614667759849472      ethnic group    yes     weak    no      yes     yes     3
782627058115641345      religion        yes     weak    no      no      yes     3
782640781290983424      ethnic group    no      no      no      no      no      0
782686657732640768      religion        yes     strong  no      no      yes     3
782787286857494528      ethnic group    no      no      no      no      no      0
782838281444683776      ethnic group    no      no      no      no      no      0
782838442044559361      ethnic group    yes     weak    weak    no      no      1
782861476126162944      religion        no      no      no      no      no      0
```

Annotated corpora for SA et. al. Evaluation

- Evaluation of the annotated data:
    - by comparing the results produced by the human annotators and calculating their disagreement
    - by training systems and then comparing their results with the data annotated by humans
- Annotation schemes: standards?
    - Evaluation campaigns and shared tasks

      Semeval (mostly English)

      Evalita (Italian)

      Ibereval (Spanish)

      …

Inter-annotator agreement (IAA)

- Rigorous methodologies for measuring the inter-annotator agreement
  - Cohen’s kappa-like measures (two coders)
  - Fleiss’s kappa measure (generalization to more than two coders)
  - http://www.aclweb.org/anthology/J08-4004: Inter-Coder Agreement for Computational Linguistics by Artstein & Poesio.
  - Increasing the number of annotators is the best strategy, because it reduces the chances of accidental personal biases.
  - Scales for the Interpretation of Kappa

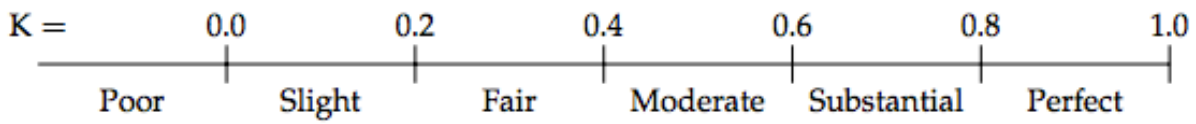
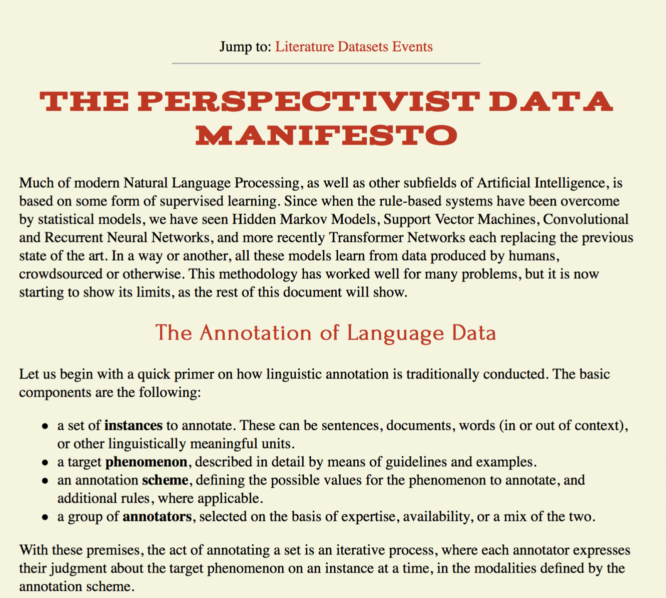


Figure 1
Kappa values and strength of agreement according to Landis and Koch (1977).

- Gold standard: manually annotated corpora

Gold standard

- Gold standard: manually annotated corpora
  - Aggregation: majority vote
  - New frontiers:
    - Perspectivist manifesto: https://pdai.info/



Jump to: Literature Datasets Events

THE PERSPECTIVIST DATA MANIFESTO

Much of modern Natural Language Processing, as well as other subfields of Artificial Intelligence, is based on some form of supervised learning. Since when the rule-based systems have been overcome by statistical models, we have seen Hidden Markov Models, Support Vector Machines, Convolutional and Recurrent Neural Networks, and more recently Transformer Networks each replacing the previous state of the art. In a way or another, all these models learn from data produced by humans, crowdsourced or otherwise. This methodology has worked well for many problems, but it is now starting to show its limits, as the rest of this document will show.

The Annotation of Language Data

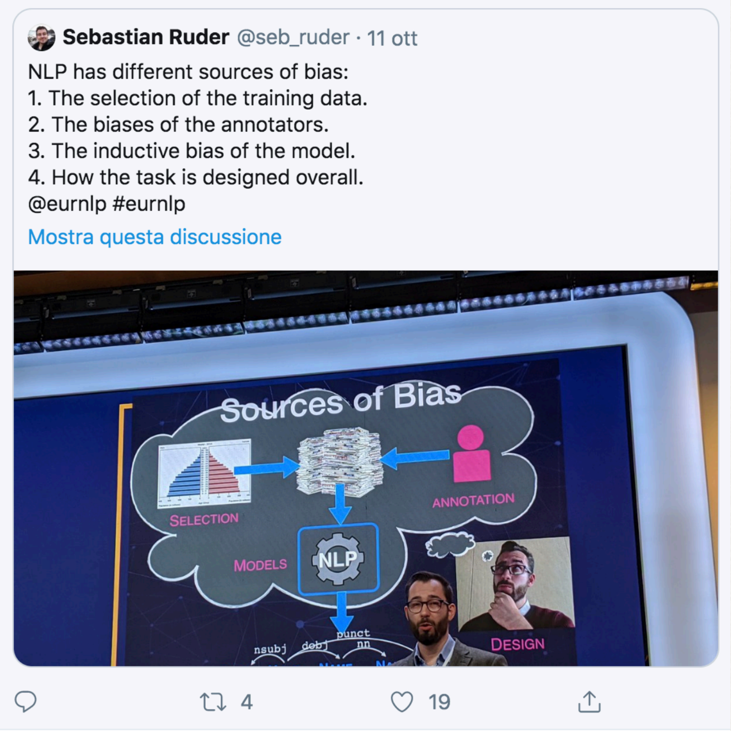
Let us begin with a quick primer on how linguistic annotation is traditionally conducted. The basic components are the following:

- a set of **instances** to annotate. These can be sentences, documents, words (in or out of context), or other linguistically meaningful units.
- a target **phenomenon**, described in detail by means of guidelines and examples.
- an annotation **scheme**, defining the possible values for the phenomenon to annotate, and additional rules, where applicable.
- a group of **annotators**, selected on the basis of expertise, availability, or a mix of the two.

With these premises, the act of annotating a set is an iterative process, where each annotator expresses their judgment about the target phenomenon on an instance at a time, in the modalities defined by the annotation scheme.

Bias in AI e NLP? We need diversity!

- Bias in developing resources and annotated corpora
  to be used as training and testing data
    - Definition of the phenomena we want to model (e.g. hate speech)
    - Selection of training data (source, authors,…)
    - Biases of the annotators
        - We need to deal with human diversity!
        - Perspective of the victims
- Machine learning with a point of view?
- Perspectivist manifesto: https://pdai.info/
- Demographic information
- Biases in selecting vulnerable groups

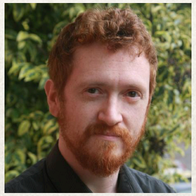


Sebastian Ruder @seb_ruder · 11 ott

NLP has different sources of bias:
1. The selection of the training data.
2. The biases of the annotators.
3. The inductive bias of the model.
4. How the task is designed overall.
@eurnlp #eurnlp

Mostra questa discussione

Recognising abuse requires expert eyes



Annotation platforms

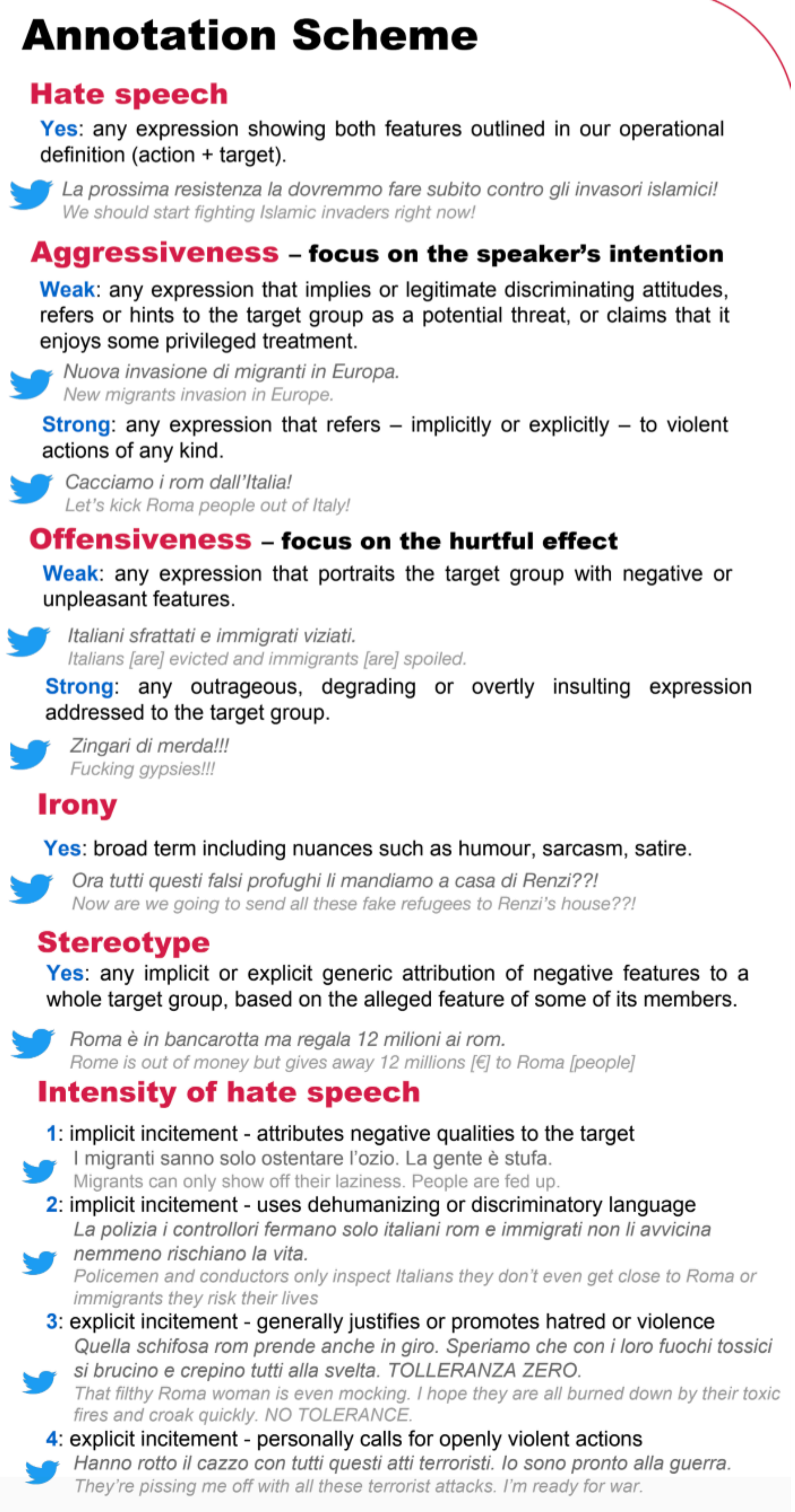
What's better from a perspectivist point of view?

- Amazon Mechanical Turks: https://www.mturk.com/
- Prolific: https://www.prolific.com/
- Appen (ex Crowdflower): https://www.appen.com/
- Label studio: https://labelstud.io/
- Home made platforms
- …

Hate Speech Corpus

- twita (http://twita.di.unito.it/) is a collection of texts from Twitter in Italian language that is continuously going on since 2012
- Hate target: immigrants
- Smaller datasets extracted from the main collection TWITA and filtered according to set of carefully selected keywords representing hate speech against migrants
    - An annotation scheme was designed for making explicit the main features of hate speech: stereotypes, aggressive attitude…
- HS as a complex and multi-layered concept
    - Multilayered annotation scheme
- Teams of annotators for applying the annotation on the datasets
- Crowdsourcing experiments for enlarging the datasets and collecting opinions of several people about what hate speech is

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, Marco Stranisci. An Italian Twitter Corpus of Hate Speech against Immigrants. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7-12, 2018.



Annotation Scheme

Hate speech

Yes: any expression showing both features outlined in our operational definition (action + target).

La prossima resistenza la dovremmo fare subito contro gli invasori islamici!
We should start fighting Islamic invaders right now!

Aggressiveness – focus on the speaker's intention

Weak: any expression that implies or legitimate discriminating attitudes, refers or hints to the target group as a potential threat, or claims that it enjoys some privileged treatment.

Nuova invasione di migranti in Europa.
New migrants invasion in Europe.

Strong: any expression that refers – implicitly or explicitly – to violent actions of any kind.

Cacciamo i rom dall'Italia!
Let's kick Roma people out of Italy!

Offensiveness – focus on the hurtful effect

Weak: any expression that portraits the target group with negative or unpleasant features.

Italiani sfrattati e immigrati viziati.
Italians [are] evicted and immigrants [are] spoiled.

Strong: any outrageous, degrading or overtly insulting expression addressed to the target group.

Zingari di merda!!!
Fucking gypsies!!!

Irony

Yes: broad term including nuances such as humour, sarcasm, satire.

Ora tutti questi falsi profughi li mandiamo a casa di Renzi??!
Now are we going to send all these fake refugees to Renzi's house??!

Stereotype

Yes: any implicit or explicit generic attribution of negative features to a whole target group, based on the alleged feature of some of its members.

Roma è in bancarotta ma regala 12 milioni ai rom.
Rome is out of money but gives away 12 millions [€] to Roma [people]

Intensity of hate speech

1: implicit incitement - attributes negative qualities to the target
I migranti sanno solo ostentare l'ozio. La gente è stufa.
Migrants can only show off their laziness. People are fed up.

2: implicit incitement - uses dehumanizing or discriminatory language
La polizia i controllori fermano solo italiani rom e immigrati non li avvicina nemmeno rischiano la vita.
Policemen and conductors only inspect Italians they don't even get close to Roma or immigrants they risk their lives

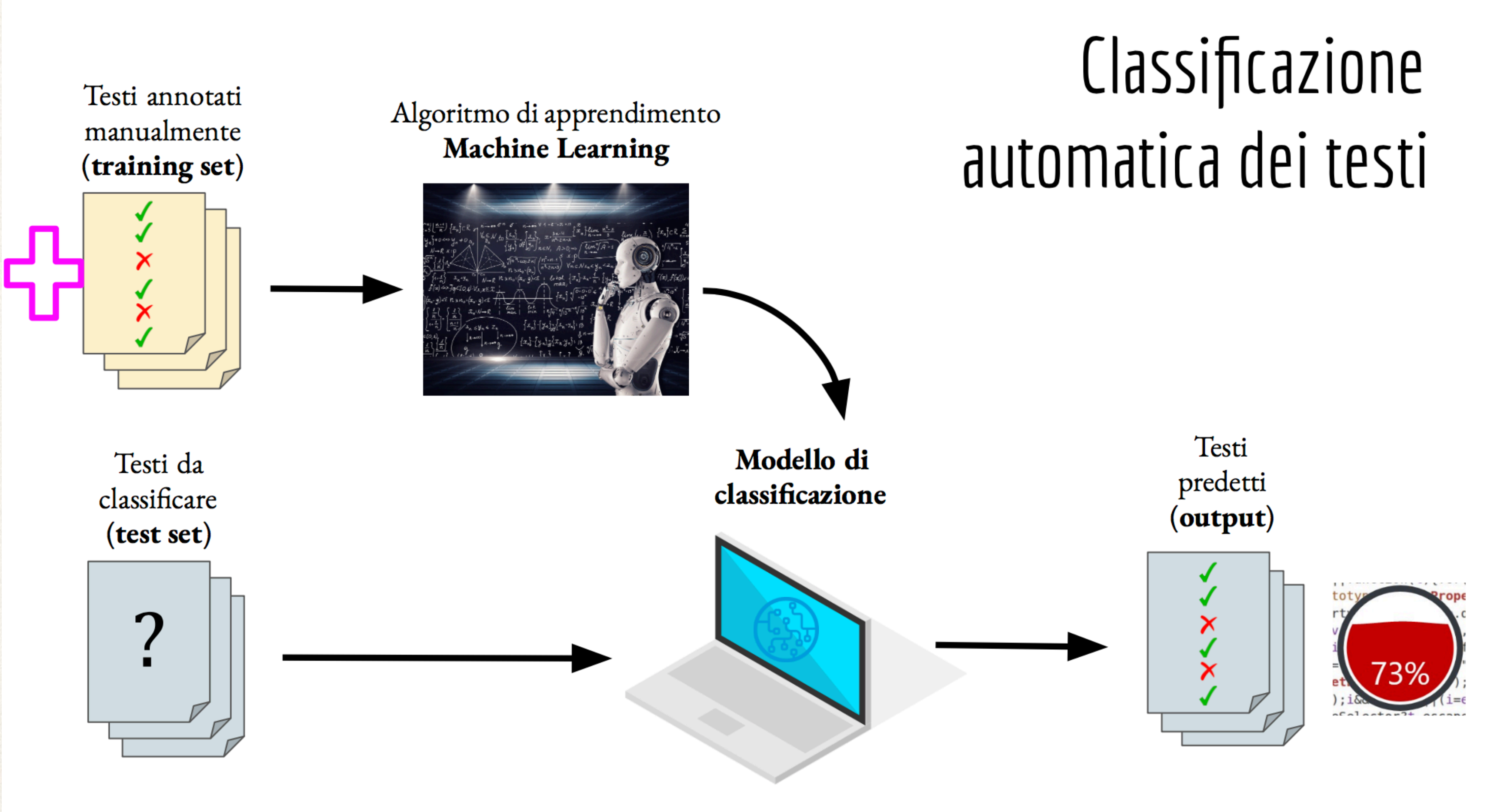
3: explicit incitement - generally justifies or promotes hatred or violence
Quella schifosa rom prende anche in giro. Speriamo che con i loro fuochi tossici si brucino e crepino tutti alla svelta. TOLLERANZA ZERO.
That filthy Roma woman is even mocking. I hope they are all burned down by their toxic fires and croak quickly. NO TOLERANCE.

4: explicit incitement - personally calls for openly violent actions
Hanno rotto il cazzo con tutti questi atti terroristi. Io sono pronto alla guerra.
They're pissing me off with all these terrorist attacks. I'm ready for war.

Annotated corpora for SA

- Supervised text classification
  - Split: Training set, test set
  - Training: text + labels (examples of correct classifications) → model
    - Finding patterns, regularities, features!
  - Prediction: text + model → labeled text
    - Testing the system with NEW examples: comparing the results with the data annotated by humans
- Corpora for sentiment analysis are currently used for testing systems for sentiment analysis:
  - the corpus without annotation is given to be processed and annotated by the system that must be tested
  - then the result produced by the system (the annotated corpus) is compared with the annotated corpus

Test set, training set



Classificazione automatica dei testi

Testi annotati manualmente (**training set**)

Algoritmo di apprendimento **Machine Learning**

Testi da classificare (**test set**)

Modello di classificazione

Testi predetti (**output**)

Evaluation and error analysis

- Sentiment analysis systems can make mistakes:
  - to recognise an opinion which is not expressed ("false positives")
  - to not recognize an opinion present in the text ("false negative")
  - to assign a wrong polarity to an opinion (e.g. in presence of figurative language this is frequent)
  - to not understand what/who is the opinion's target
  - **Error analysis!**