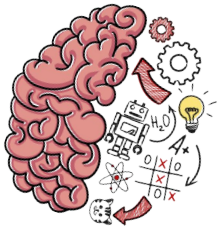


di.unito.it

DIPARTIMENTO DI INFORMATICA

Commission for Orientation and Informatics in schools

#DeactivHate

Artificial Intelligence and Computational Linguistics to counter the spread of hateful messages online

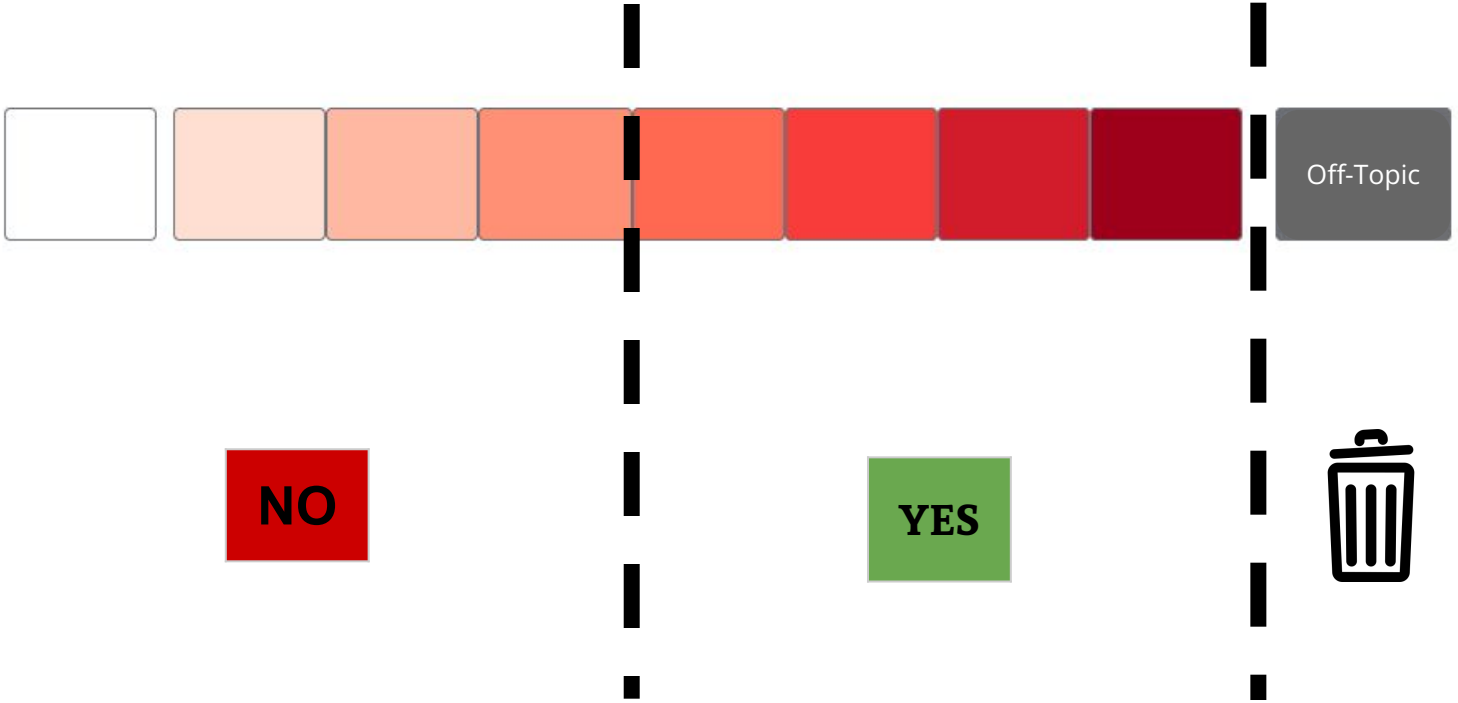
From annotation to Machine Learning

From annotation to Machine Learning

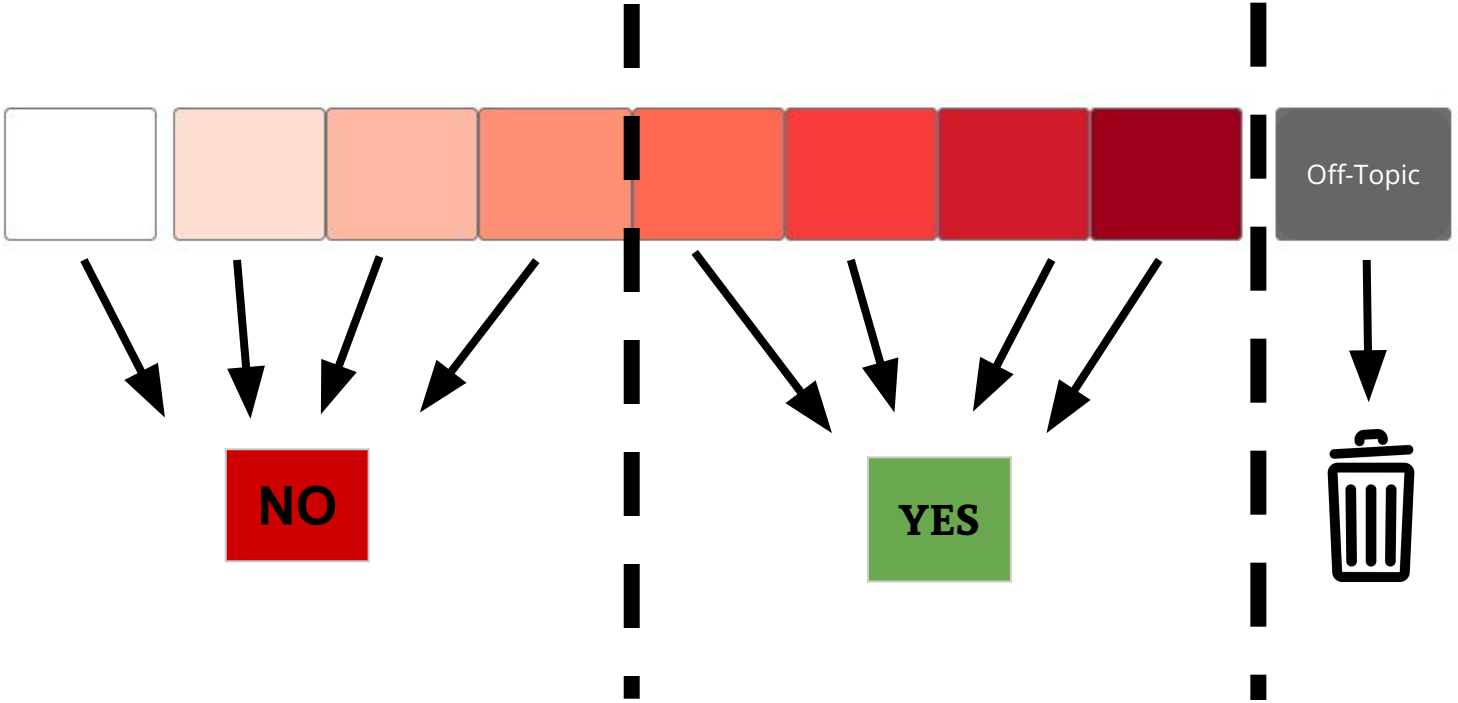
NO

YES

From annotation to Machine Learning



From annotation to Machine Learning



Machine Learning

The two main categories of machine learning

1. **Supervised learning**
2. Unsupervised learning

Machine Learning

Supervised learning

It occurs by providing the algorithm with input data called training sets, corresponding to a selection of linguistic objects that it intends to analyze automatically; the input data is prepared by expert people through an annotation process, which is necessary to train the system to learn how to classify the data.

Machine Learning

From: Profolan	Object: The best remedy for baldness
From: 7Slim	Object: How to lose weight?
From: School.edu	Object: Announcement from the principal
From: School.edu	Object: Suspension of lessons

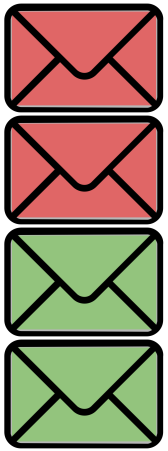
Machine Learning

From: Profolan Object: The best remedy for baldness

From: 7Slim Object: How to lose weight?

From: School.edu Object: Announcement from the principal

From: School.edu Object: Suspension of lessons



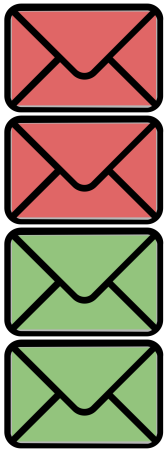
Machine Learning

From: Profolan Object: The best remedy for baldness

From: 7Slim Object: How to lose weight?

From: School.edu Object: Announcement from the principal

From: School.edu Object: Suspension of lessons



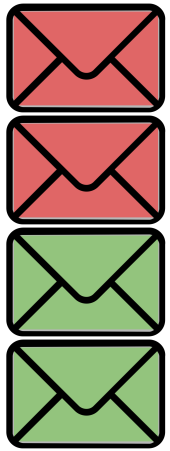
Machine Learning

From: Profolan Object: The best remedy for baldness

From: 7Slim Object: How to lose weight?

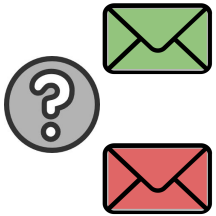
From: School.edu Object: Announcement from the principal

From: School.edu Object: Suspension of lessons



Spam or **E-mail**?

From: Nicoin Object: The easiest way to quit smoking



Machine Learning

The two main categories of machine learning

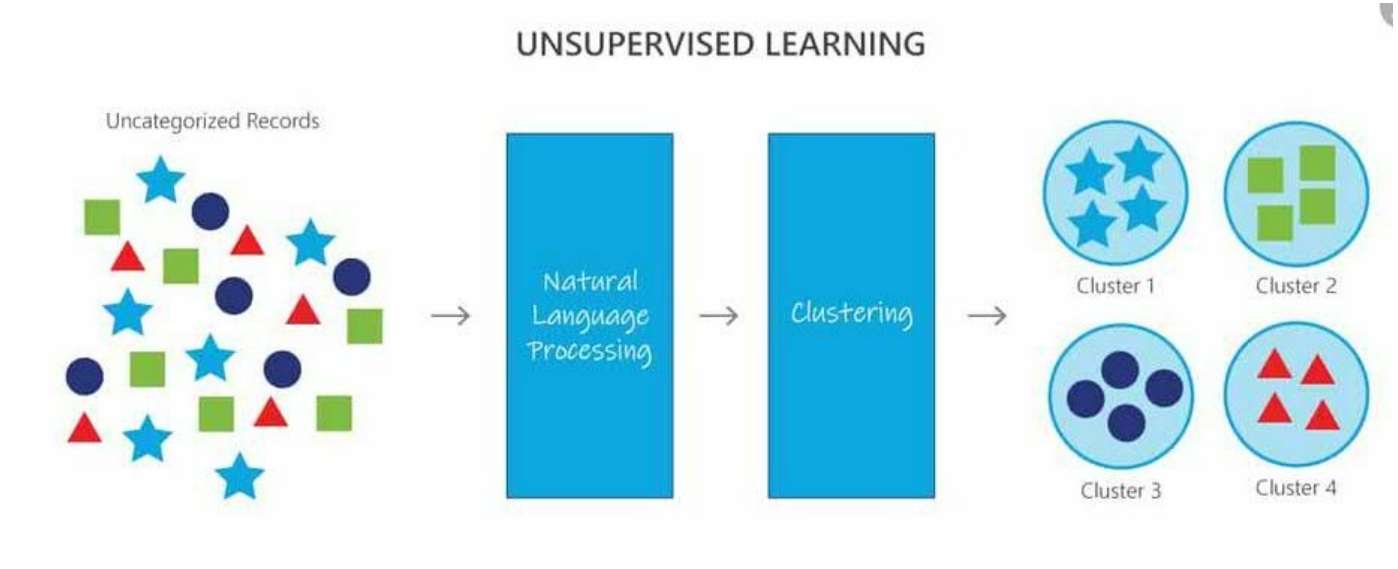
1. Supervised learning
2. **Unsupervised learning**

Machine Learning Methods

Unsupervised learning

It occurs without the use of training data prepared by expert people. The model tries to independently extract input data without having rules decided *a priori*.

Machine Learning Methods



Some useful vocabulary

1. **Features or attributes:** i.e. linguistic variables that represent the input data, chosen by the expert because they are considered relevant for the learning task (in the examples we will see today the object is a text)
2. **Task:** or linguistic task, it is the problem you want to address and solve (for example, the classification of emails between 'Spam' and 'Regular emails')
3. **Learning algorithm:** is a mathematical/statistical/logical method used to create a linguistic model of text classification
4. **Training data:** it is the data that is used by a learning algorithm to create a model
5. **The model:** it is the output of a learning algorithm obtained by training it with a given set of training data

Some useful vocabulary

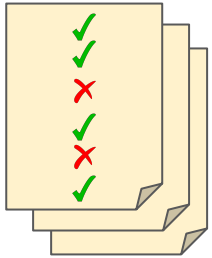
The training data (**train set**) constitutes a dataset of texts classified (usually by manual annotators) distinguishing between two or more classes.

The input of a text classification task is a set of texts (**test set**) that we want to classify by discerning between two or more classes (the same classes used in the training set).

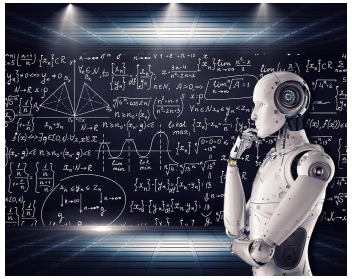
The output is an automatic prediction of the class to which each element of the test set belongs.

Machine Learning workflow

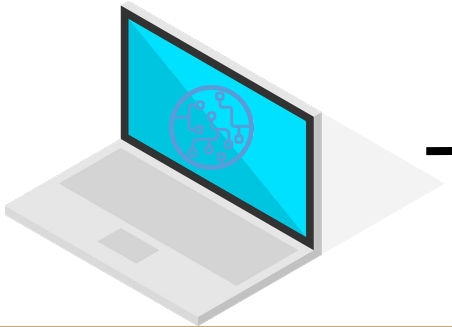
Manually annotated texts (**training set**)



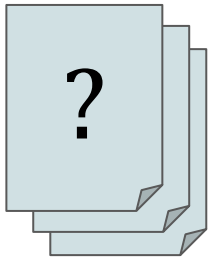
Learning algorithm **Machine Learning**



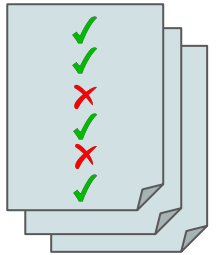
Classification model



Texts to classify (**test set**)

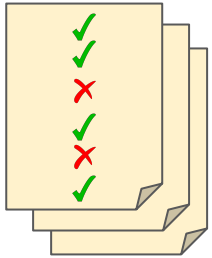


Predicted texts (**output**)



Machine Learning workflow

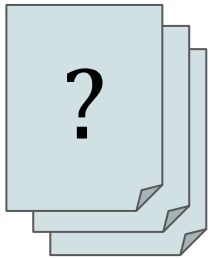
Manually annotated texts (**training set**)



=

1047919240848838656	@user @user This is an example of text extracted from social media, and it contains hateful expressions HATEFUL HATEFUL #HATE	**YES**
...	....	...
1055101652557094913	@user @user This is an example of text extracted from social media, and it contains RAINBOWS AND STARS and colors and cute kitties	**NO**

Texts to classify (**test set**)

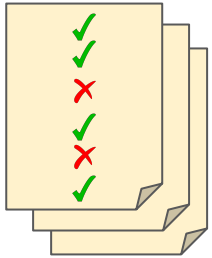


=

1047919240821123465	@user This is an example of text extracted from social media, and we don't know how to classify it	?
...	....	
1234532454568894345	@user This is a second example of text extracted from social media, and we don't know how to classify it	?

Machine Learning workflow

Manually annotated texts (**training set**)



=

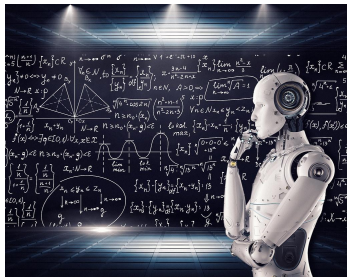
1047919240848838656	@user @user This is an example of text extracted from social media, and it contains hateful expressions HATEFUL HATEFUL #HATE	**YES**
...	....	...
1055101652557094913	@user @user This is an example of text extracted from social media, and it contains RAINBOWS AND STARS and colors and cute kitties	**NO**

1	1	0	...	1	0	**YES**
...	...	...	...	...	...	...
0	0	1	...	0	1	**NO**

Machine Learning workflow

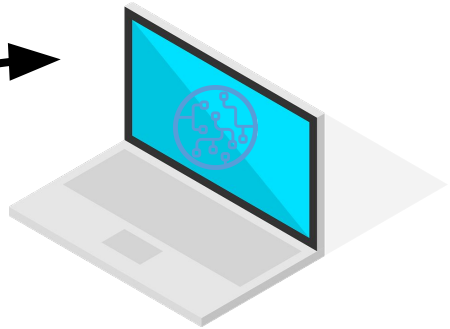
Training set texts
(**vectorized**)

1	1	0	...	1	0	**YES**
...	...	...	...	...	...	...
0	0	1	...	0	1	**NO**

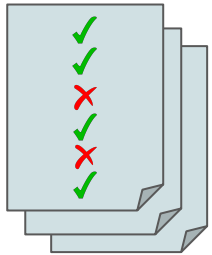


Learning algorithm
Machine Learning

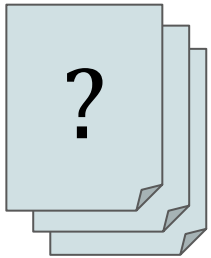
Classification model



Predicted texts
(**output**)



Machine Learning workflow



Texts to classify (**test set**)

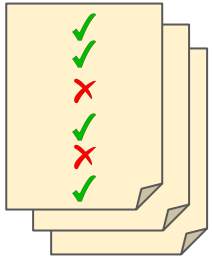
=

1047919240821123465	@user This is an example of text extracted from social media, and we don't know how to classify it	?
...	....	
1234532454568894345	@user This is a second example of text extracted from social media, and we don't know how to classify it	?

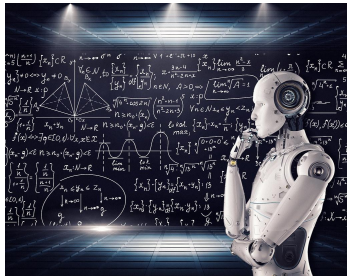
1047919240821123465	@user This is an example of text extracted from social media, and now we know how to classify it	**YES**
...	....	...
1234532454568894345	@user This is a second example of text extracted from social media, and now we know how to classify it	**NO**

Machine Learning workflow

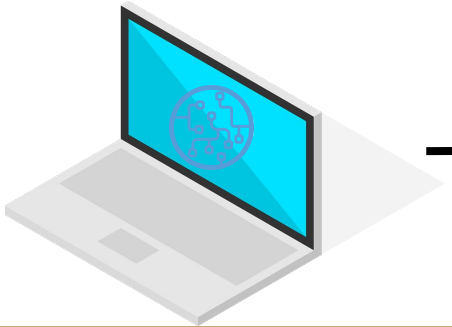
Manually annotated texts (**training set**)



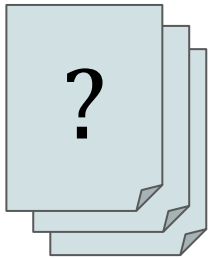
Learning algorithm **Machine Learning**



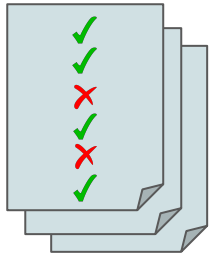
Classification model



Texts to classify (**test set**)



Predicted texts (**output**)



Automatic text classification

What we need to do:

1. Define a task to solve
2. Collect a dataset of texts
3. Create a training set of annotated texts and a test set
4. Represent texts using features
5. Use a machine learning algorithm
6. Create the NLP model
7. Evaluate the model
8. Automatically label new texts you've never seen before

1. Define a task to solve

Hate Speech Detection

Hate speech consists of a specific form of discrimination that is expressed not through actions or omissions, but through deplorable mode of manifestation of thought. Disseminated and reiterated through the Internet, these forms of expression have the effect of fueling prejudices, consolidating stereotypes and reinforcing the hostility of certain groups of people, usually in the majority or in a dominant position in a given social context, towards others groups with different characteristics, generally minorities.
https://www.agendadigitale.eu/cultura-digitale/le-strategie-di-contrasto-allodio-online-nellunione-europea-46113/

2.1. Collect a dataset of texts

Twitter is the social media most used by us researchers for social media analysis because it provides a series of APIs (Application Programming Interface) that allow us to collect public posts published by users of this platform (read more https://apps.twitter.com/)

There are libraries for the most popular programming languages that allow you to easily use the API, and a lot of ready code is made freely available by programmers.

2.2. Collect a dataset of texts

We could query the Twitter API asking to receive only tweets that contain certain keywords.

For example, we could focus on the analysis of online hate speech aimed at immigrants and analyze tweets that contain at least one of these words:

[migrant, immigrant, clandestine, foreigner]

2.3. Collect a dataset of texts

We format the data in the way that is easiest for us. We could, for example, create a CSV file in which, in each line, there is the id of a tweet (its unique code) and its textual content.

id	testo
1046604997545414656	New, yet another, demonstration for #illegal #immigrants in #Milan. Radical chic, do-gooders, immigrationists, trade unions, co-ops, non-profit organisations, NGOs, Anpi with the queen of anti-Italians, Boldrini. As usual they are only interested in the hospitality business. https://t.co/N2qiaDh64Y
1046620978544095233	Milan, thousands in the streets against racism and intolerance. Anpi: “The real threat is not migrants but mafias” https://t.co/YutiphXCrl via @fattoquotidiano
1046639560581685250	@user @user thank goodness there are THEM demonstrating for "more MIGRANTS and more BAKERS for ALL..."

3.1. Collect a dataset of texts

We need to define which classes (**labels**) the tweets can belong to.

We could opt for a binary classification (remember Spam and E-mail?).

The tweet can contain hate speech (I use the YES label) or not (I use the NO label).

3.2. Create a training set

It is absolutely not easy to decide whether a tweet contains hate speech or not. It often depends on the sensitivity of the person reading it.

1. We need to provide annotators with **guidelines** that define what hate speech is for the purposes of our task

2. You need to ask multiple annotators to annotate the same tweet in order to have as many opinions as possible

3.5. Create a training set

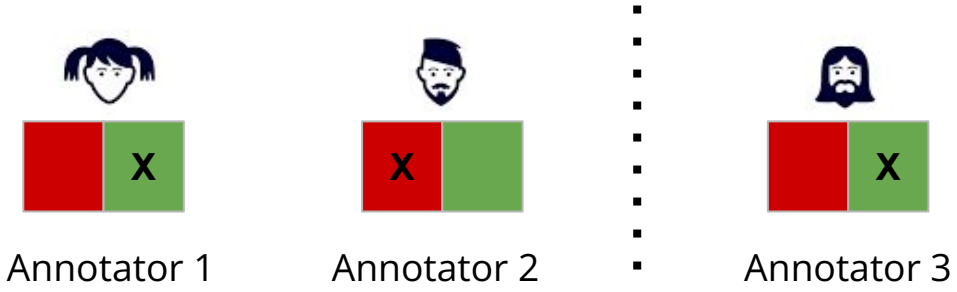
If the manual annotators label the same tweet differently, you can choose to ask for a further opinion (another annotation) and select the most used label.

1046604997545414656	New, yet another, demonstration for #illegal #immigrants in #Milan. Radical chic, do-gooders, immigrationists, trade unions, co-ops, non-profit organisations, NGOs, Anpi with the queen of anti-Italians, Boldrini. As usual they are only interested in the hospitality business. https://t.co/N2qiaDh64Y	?

3.5. Create a training set

If the manual annotators label the same tweet differently, you can choose to ask for a further opinion (another annotation) and select the most used label.

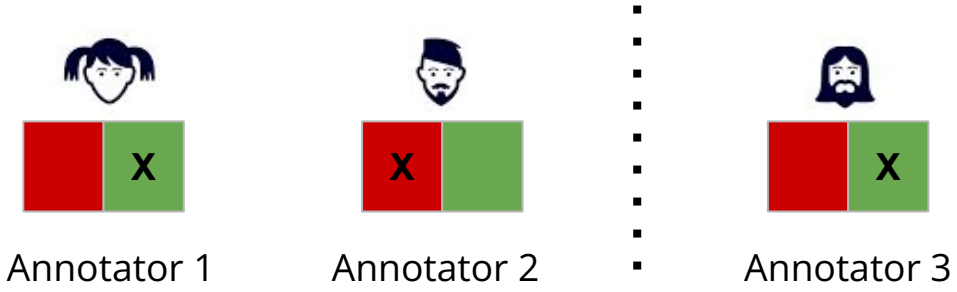
1046604997545414656	New, yet another, demonstration for #illegal #immigrants in #Milan. Radical chic, do-gooders, immigrationists, trade unions, co-ops, non-profit organisations, NGOs, Anpi with the queen of anti-Italians, Boldrini. As usual they are only interested in the hospitality business. https://t.co/N2qiaDh64Y	?



3.5. Create a training set

If the manual annotators label the same tweet differently, you can choose to ask for a further opinion (another annotation) and select the most used label.

1046604997545414656	New, yet another, demonstration for #illegal #immigrants in #Milan. Radical chic, do-gooders, immigrationists, trade unions, co-ops, non-profit organisations, NGOs, Anpi with the queen of anti-Italians, Boldrini. As usual they are only interested in the hospitality business. https://t.co/N2qiaDh64Y	**YES**



3.7. Create a training set

1047919240848838656	@user @user Let's take back our city cleansed of illegal immigrants, illegal immigrants, criminals, migrants, drug dealers and Nigerians and gypsies	**YES**
1048255760575152129	You have to give it only to Italians! How disgusting these foreigners always want everything for free! @user don't be so stupid as to give it to everyone except Italians born in Italy, no foreigners! Read this foreign shit they just want easy money!👇https://t.co/zanyD7Y0kL	**YES**
1053063253654409216	Not building cycle paths to spite migrants is the racist version of cutting it off to spite your wife. https://t.co/zzovf3WwK2	**NO**
1055101652557094913	"Jackal!" those who go on a cruise to film immigrants shout at the Minister of the Interior, who has rushed to investigate a murder.🤣🤣 #desiree	**NO**

Automatic text classification

What we need to do:

1. ~~Define a task to solve~~
2. ~~Collect a dataset of texts~~
3. ~~Create a training set of annotated texts and a test set~~
4. Represent texts using features
5. Use a machine learning algorithm
6. Create the NLP model
7. Evaluate the model
8. Automatically label new texts you've never seen before