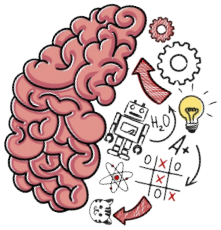


di.unito.it

DIPARTIMENTO DI INFORMATICA

Commissione Orientamento e Informatica nelle scuole

#DEACTIVHATE

ARTIFICIAL INTELLIGENCE AND COMPUTATIONAL LINGUISTICS TO COUNTER THE SPREAD OF HATEFUL MESSAGES ONLINE

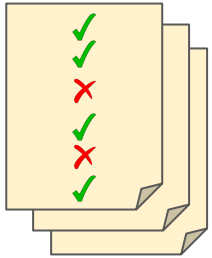
Machine Learning

The two main categories of machine learning

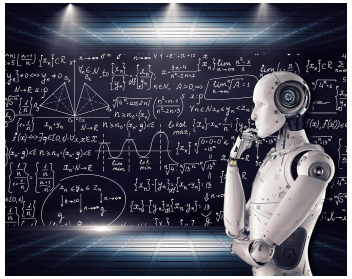
1. Supervised learning
2. Unsupervised learning

Machine Learning workflow

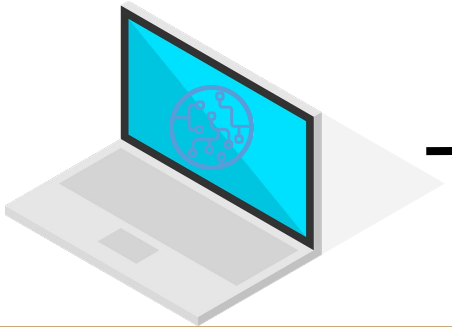
Manually annotated texts (**training set**)



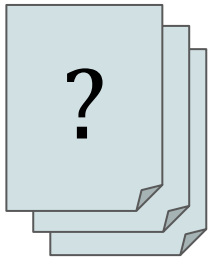
Learning algorithm **Machine Learning**



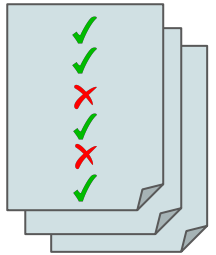
Classification model



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

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/

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


twitter

Automatic text classification

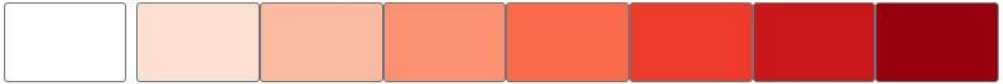
What we need to do:

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

[@utente] [@utente] non è nel vostro diritto aggredire giornalisti (come accaduto) o urlare slogan che offendono dignità professionale di categorie (da giornalisti terroristi a medici assassini).

(1/15)

Qual è il livello di hate speech nei confronti di musulmani, immigrati o rom presente in questo tweet?



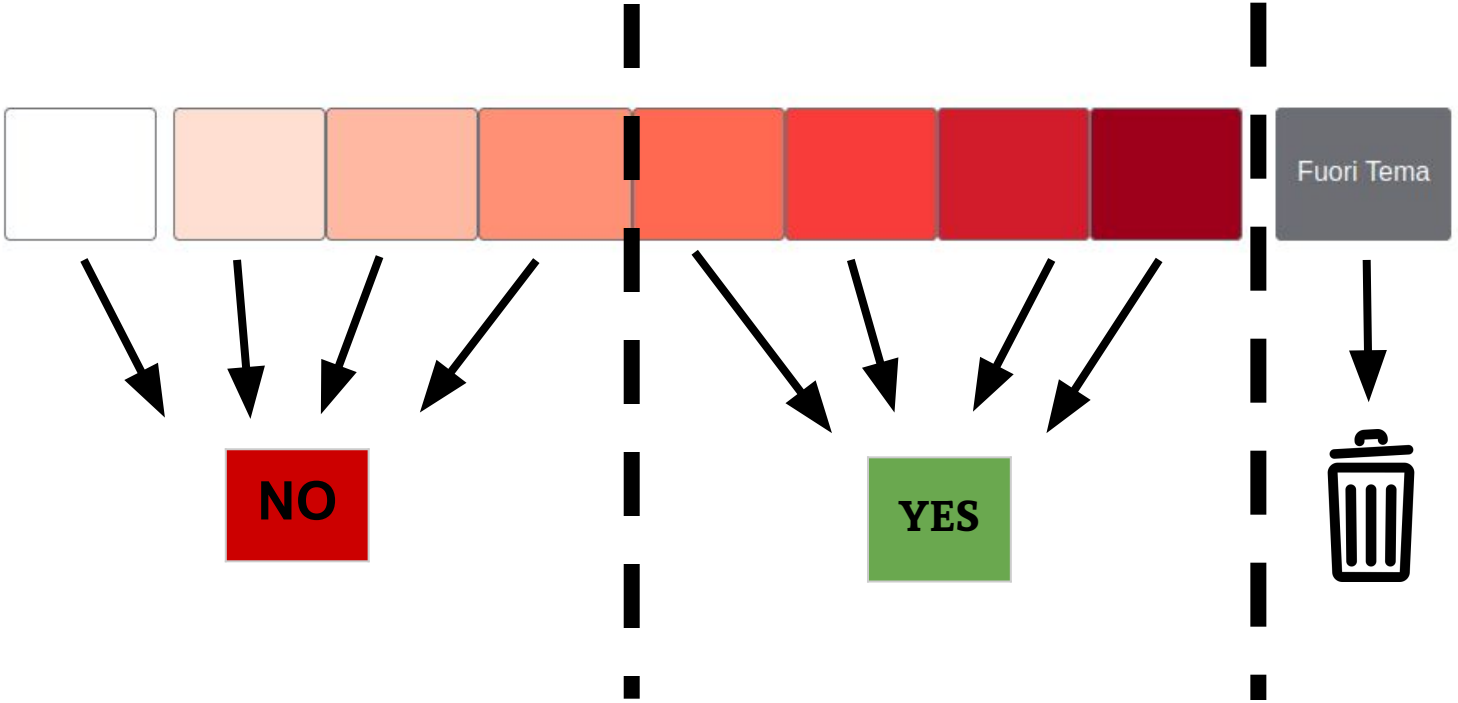
Non presente Ironia/Sarcasmo/Humor

Non presente Offensività

Non presente Stereotipo

Prosegui

From annotation to Machine Learning



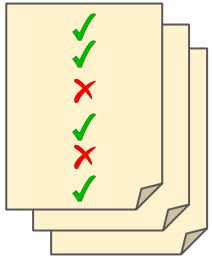
Automatic text classification

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

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**

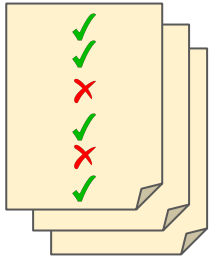
Automatic text classification

What we need to do:

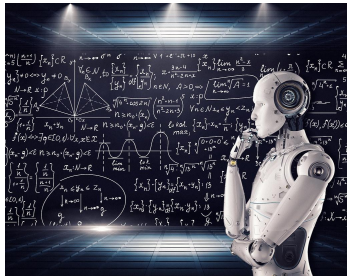
1. ~~Define a task to solve~~
2. ~~Collect a dataset of texts~~
3. ~~Create a training set of annotated texts and a test set~~
4. ~~Represent texts using features~~
5. **Use a machine learning algorithm**
6. Create the NLP model
7. Evaluate the model
8. Automatically label new texts you've never seen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

Machine Learning workflow

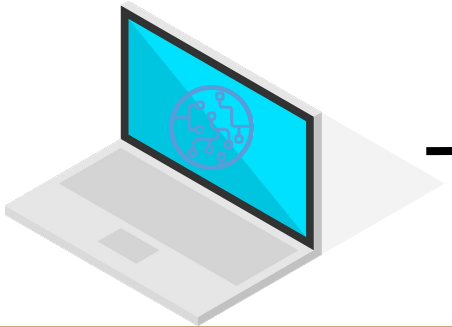
Manually annotated texts (**training set**)



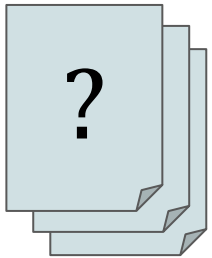
Learning algorithm **Machine Learning**



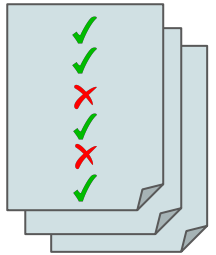
Classification model



Texts to classify (**test set**)



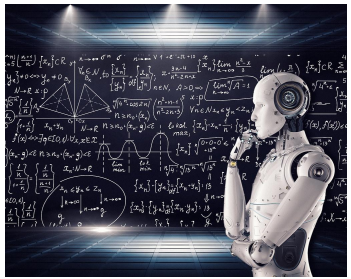
Predicted texts (**output**)



Machine Learning workflow

Training set texts
(**vectorized**)

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



Learning algorithm
Machine Learning

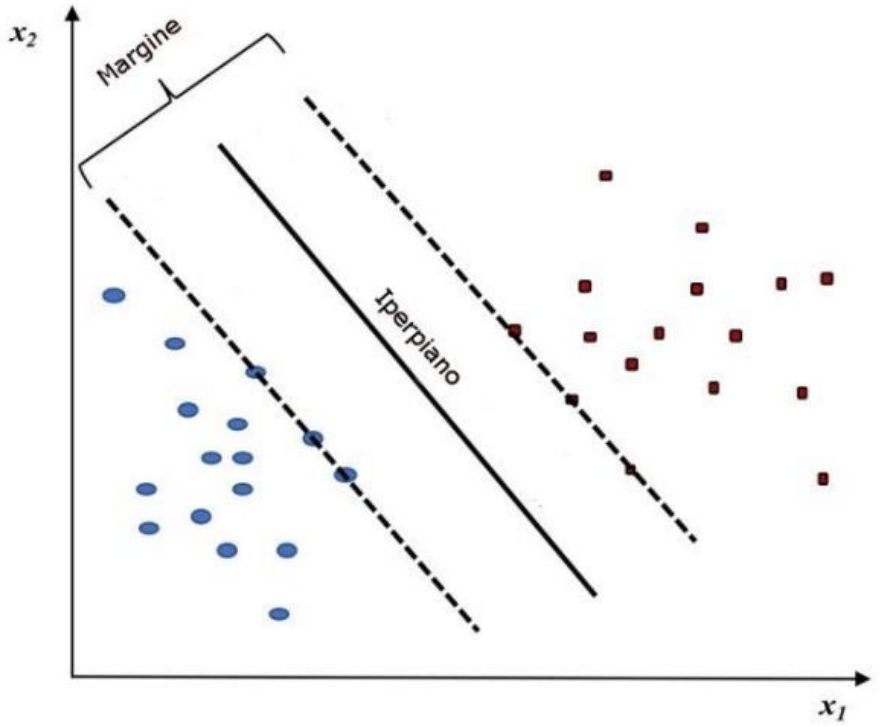
Machine Learning Algorithm

The choice of the most suitable machine learning algorithm to tackle a given task depends on many factors:

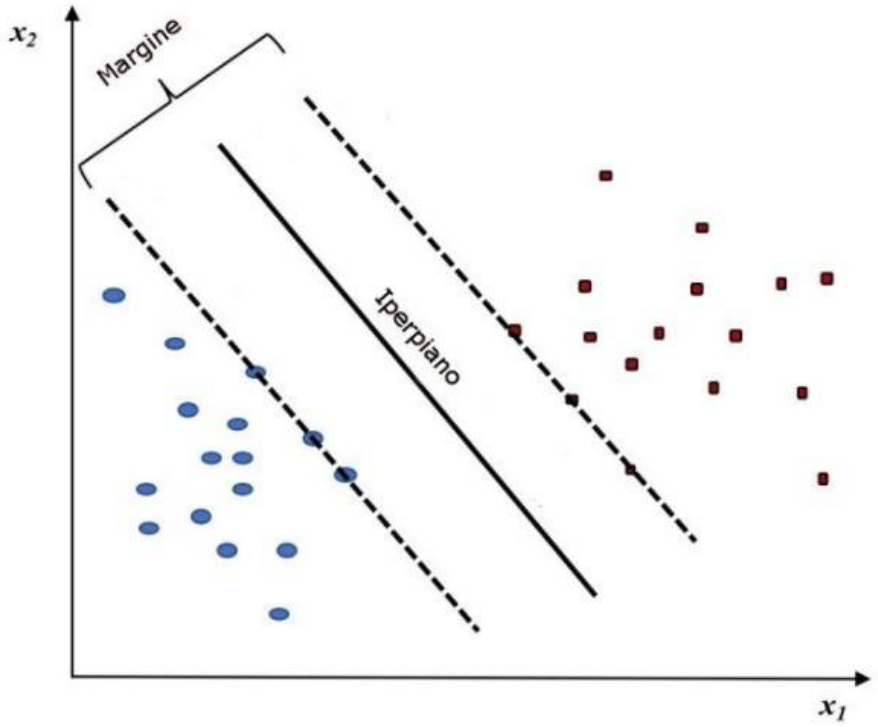
Task type, training set size, etc...

Although some factors can be evaluated before testing, evaluating which one performs best is fundamental → Evaluation

Support-Vector Machine (SVM)



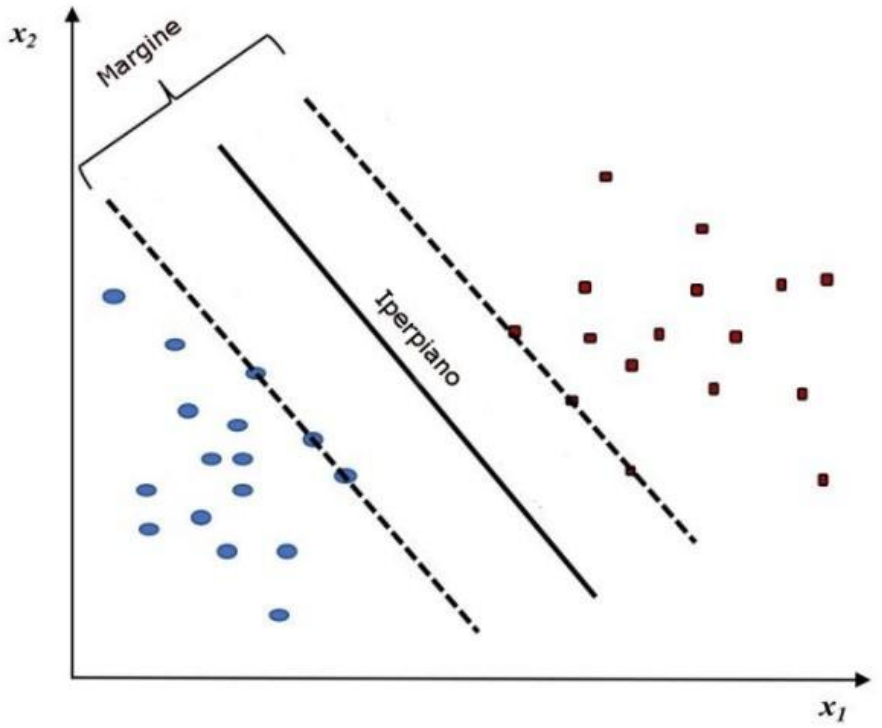
Support-Vector Machine (SVM)



Objective:

- find the hyperplane (or line in a 2D plane) that best divides a dataset into 2 classes

Support-Vector Machine (SVM)



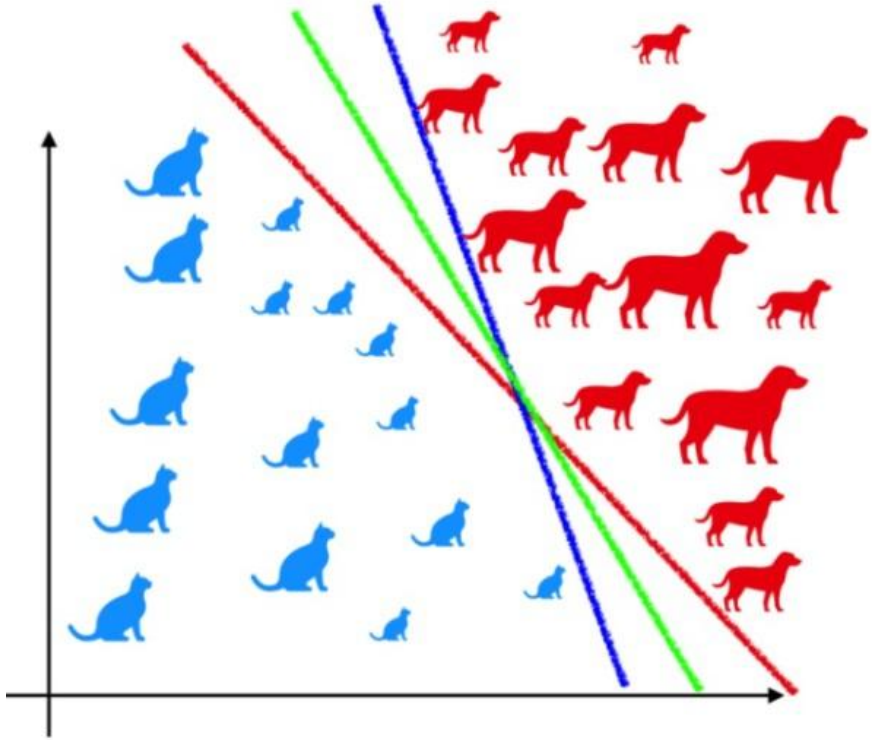
Objective:

- find the hyperplane (or line in a 2D plane) that best divides a dataset into 2 classes

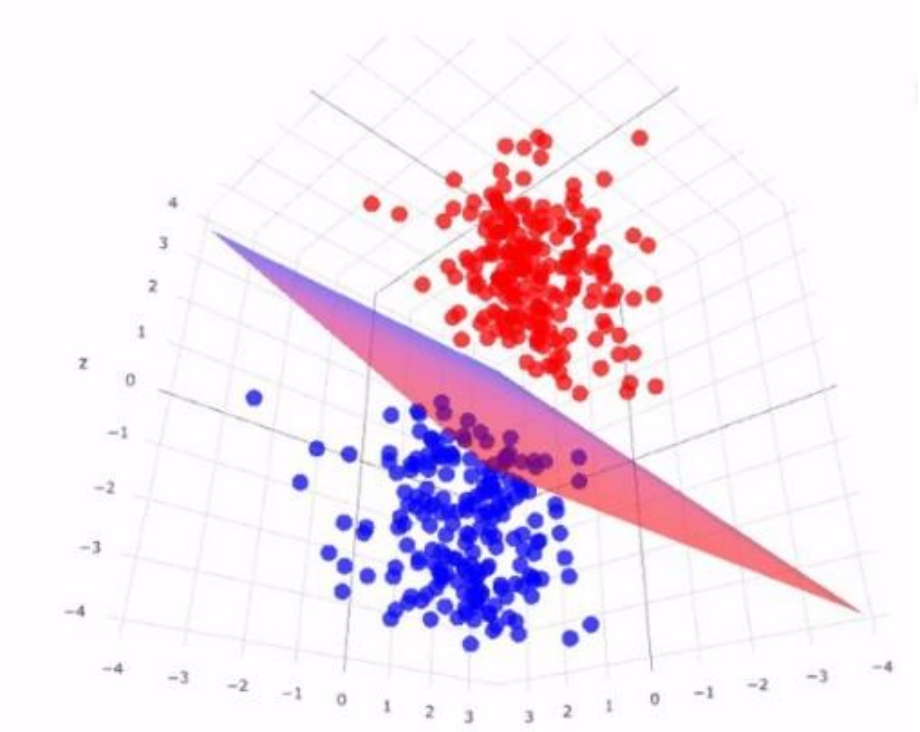
Hyperplane:

- which maximizes the distance between the closest points of the 2 classes. The greater the margin, the lower the possibility of error of the classifier.

Support-Vector Machine (SVM)



Support-Vector Machine (SVM)



Automatic text classification

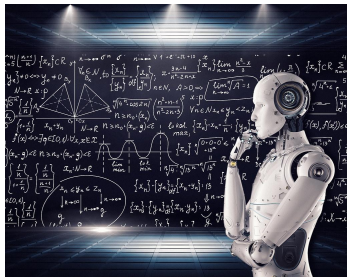
What we need to do:

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

Machine Learning workflow

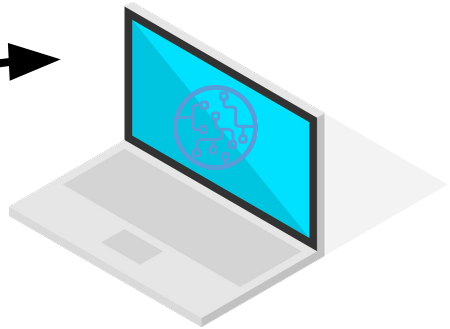
Training set texts
(**vectorized**)

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



Learning algorithm
Machine Learning

Classification model



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

Model Evaluation

The label assigned by the model is called a **prediction**.

The label assigned by the human annotator is called **gold label**.

The model must **predict** (classify) the label (class) assigned to a text by getting as close as possible to the choices made by the human annotator (the gold label).

... **But the model should not classify the texts already used during the training phase**. It would be as if during the lesson, I was given the answers to the exam questions ...

Model Evaluation

This involves evaluating
how "similar" the model prediction
is to manual annotation

Predicted label **Gold label**

vs

Two common evaluation metrics are:

- **Accuracy**
- F-measure

Model Evaluation

Accuracy measures the number of times the model classifies correctly the label

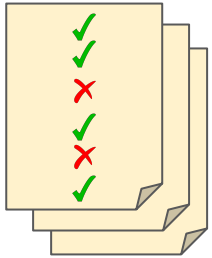
Automatic text classification

What we need to do:

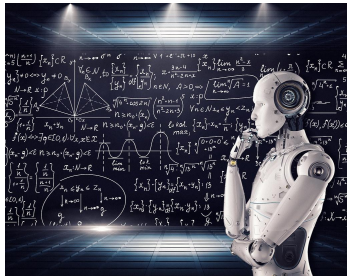
1. ~~Define a task to solve~~
2. ~~Collect a dataset of texts~~
3. ~~Create a training set of annotated texts and a test set~~
4. ~~Represent texts using features~~
5. ~~Use a machine learning algorithm~~
6. ~~Create the NLP model~~
7. ~~Evaluate the model~~
8. **Automatically label new texts you've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

Machine Learning workflow

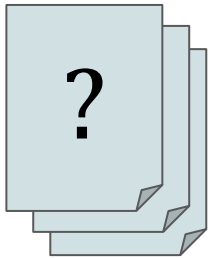
Manually annotated texts **(training set)**



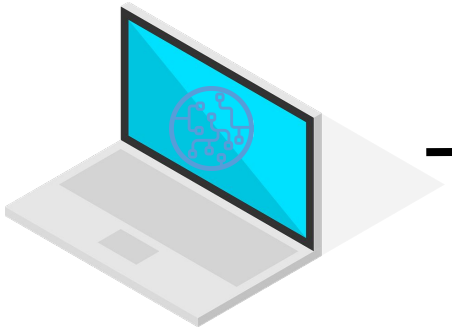
Learning algorithm **Machine Learning**



Texts to classify **(test set)**



Classification model



Predicted texts **(output)**