

# **Artificial Intelligence and/or for the Social Sciences**

Active Tigger tutorial

# Today's program

- Methodological choices in an age of textual abundance
- **An Introduction to the interface ActiveTigger**
- Going further
- Q&A

# Active Tigger

An Interface to speed up Annotation with LLMs

# Introduction

## Textual data overload

How can we make the most of analyzing massive datasets without sacrificing nuance?

- Possible with qualitative methods
- But always human limits on scale

# Introduction

## Example

Objective: quantify online abuse targeting French politicians

Pre-trained models exist, but they do not necessarily reflect types of expression used in French Twittersphere, nor the types of nuance that I wanted to capture

Une petite camomille vous ferez du bien 🤡🤡 pauvre garçon 🍵🍵

[Translate post](#)

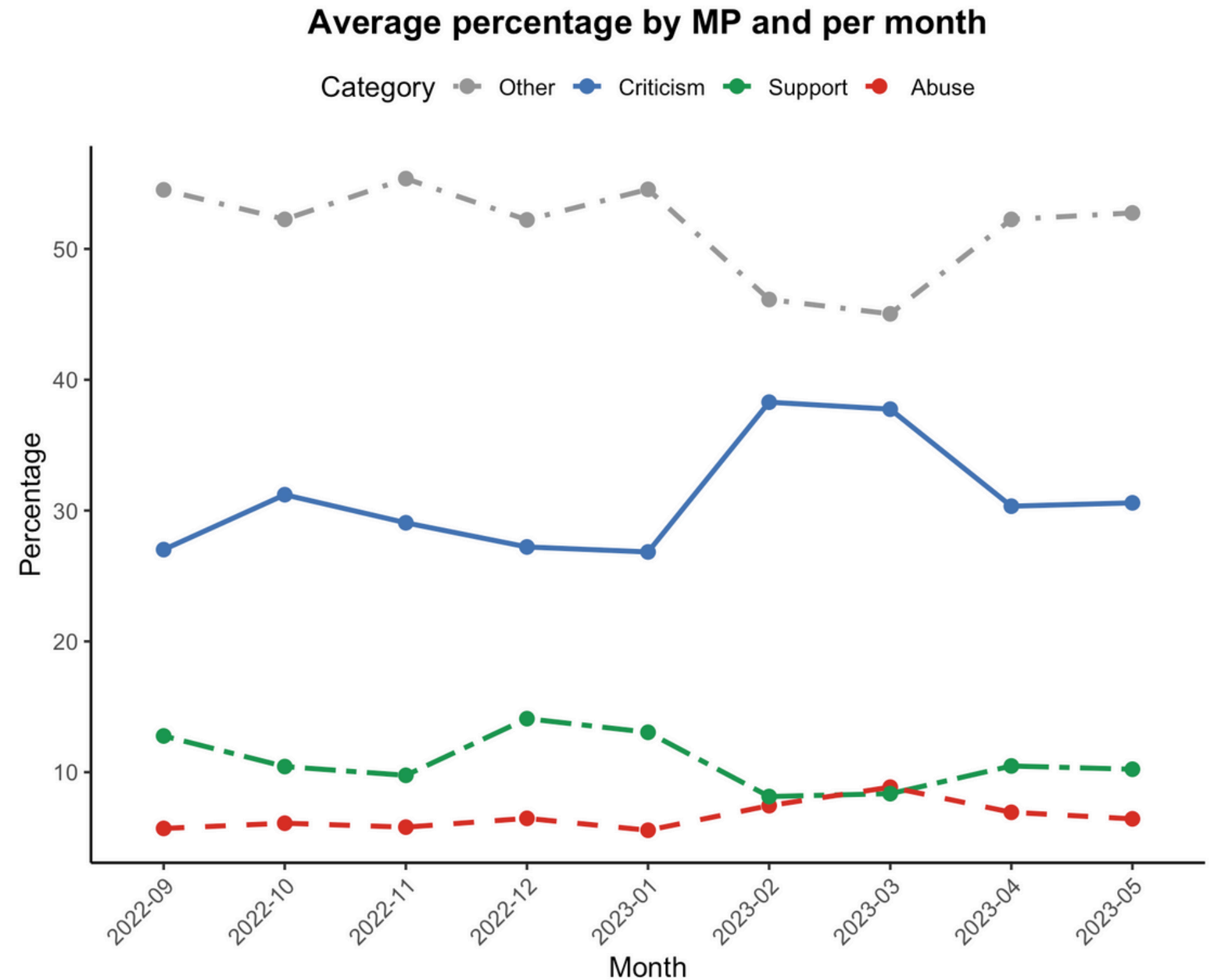
12:40 AM · Nov 16, 2022

# Introduction

## Example

Solution: train a bespoke classifier to identify online abuse according to my own coding scheme

3000 manual annotation =>  
30+ million of classified tweets



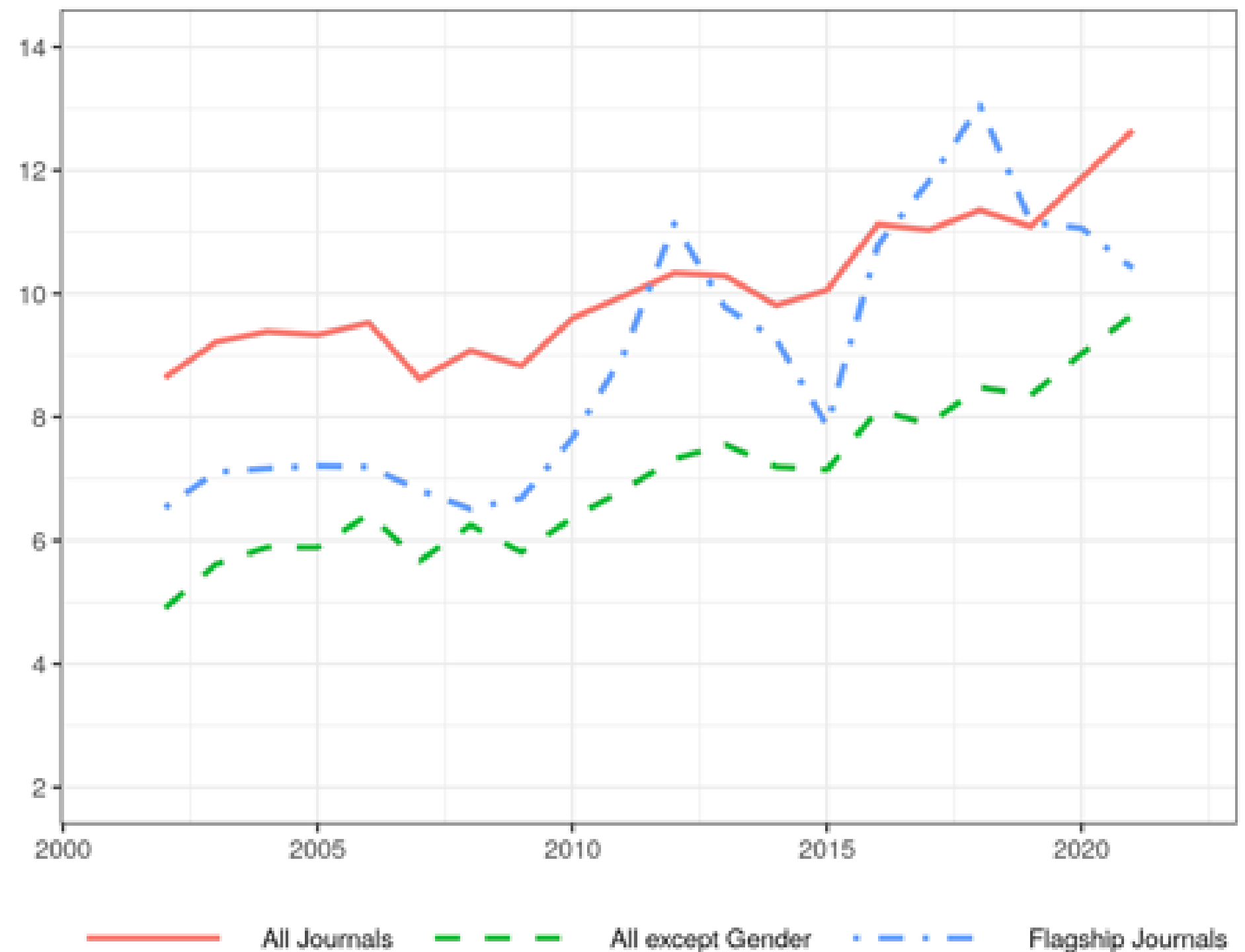
# Introduction

## Example

Is a gender perspective becoming more common in French social science?

Boelaert, Coavoux, Gollac, Ollion

Recognizing gender-related terms in article abstracts (50 000+)



# Machine learning

## Annotating data

- A standalone method
- Or a step to apply a coding scheme on a larger corpus (via a model)



# Machine learning

## Supervised learning

= training a model to recognize your coding scheme (through annotations) to then extend it to a larger dataset

# Machine learning

## Supervised learning

Recently, *large language models* (LLMs) have become powerful tools for classification, including in social sciences

It is, for example, possible to ask ChatGPT to perform such tasks (benefits & limitations)

Balance between function, time, and cost => encoder models (**BERT**).

# Machine learning

## Supervised learning

**BERT**: A language model based on neural networks + Transformers architecture  
Bidirectional Encoder Representations from Transformers

Pros:

- Takes into account the context surrounding a given word or sphrase
- Allows us to categorize texts on precise criteria

# Machine learning

(Augmented) classification

So what are the steps?

1. Defining categories (what are you looking for?)
2. Patient annotation of a smaller corpus
3. Training and validating model
4. Extension (“inferring”) on total dataset

Do, S., Ollion, É., and Shen, R (2024) The augmented social scientist

# Machine learning

## (Augmented) classification

So what are the steps?

1. Defining categories (what are you looking for?)
2. Patient annotation of a smaller corpus
3. Training and validating model
4. Extension (“inferring”) on total dataset



# Machine learning

## (Augmented) classification

Annotation requires:

1. a deep understanding of your data
2. a codebook
3. patience and iterations



# ActiveTigger

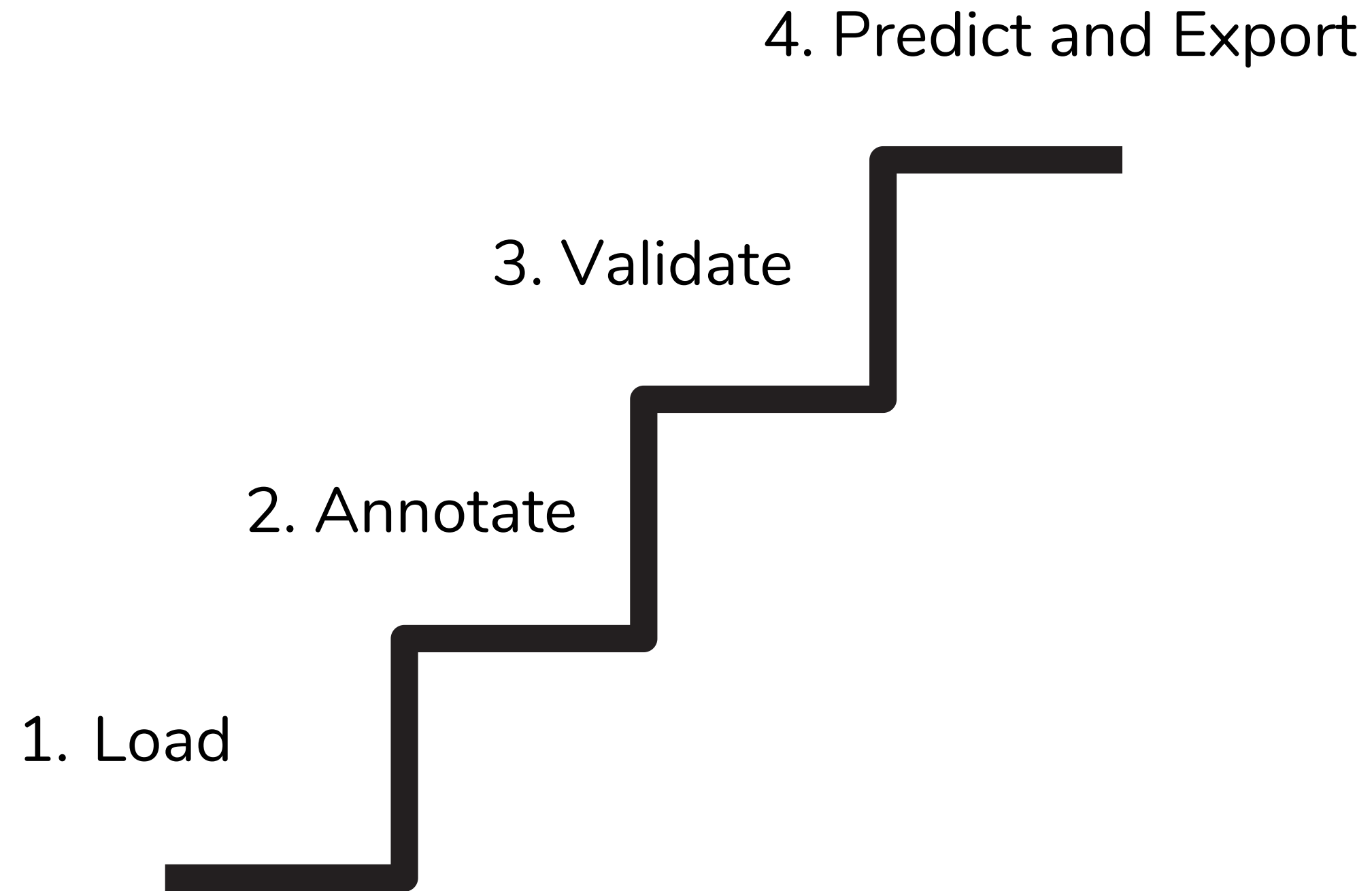
ActiveTigger

= an interface simplifying the annotation process

- A user-friendly, non-coding interface
- Offers ways to accelerate the annotation process
- Provides quality indicators
- Trains and applies model on data

# ActiveTigger

## Key Functions





# ActiveTigger

## Create a Project

You will (only) need a file (**csv, xlsx or parquet**) containing the texts of interest (unit can vary), with unique identifiers

If you want, you can separate your data into a **training dataset** and a **test dataset** (best practice)

Id column (they need to be unique)

Row number

Text columns (all the selected fields will be concatenated)

Select...

Language of the corpus (for tokenization and word segmentation)

English

Column for existing annotations (optional)

Select...

Contextual information columns (optional)

Select...

Number of elements in the train set (limit : 100.000)

100

Number of elements in the test set (not already annotated)

0


Empty testset ☐


Stratify the test set by

Select...

# ActiveTigger

## Prepare: Devise an Annotation Scheme

 [ActiveTigger](#) [Projects](#) [Account](#) [Users](#) Logged as eollion



[unattribued](#) [default](#)

[Prepare](#)

[Explore](#)

[Annotate](#)

[Train](#)

[Test](#)

[Predict](#)

[Export](#)

Process: 0

GPU:1.5 / 85.1 Go

Labels

Features

Codebook

Import annotations

Add a label

Create, delete or rename labels.

The current scheme is a **multiclass**

Available labels

unattributed

▼

🗑️

New label

Enter new label

+

Convert label

Replace selected label to the new one 🔍

# ActiveTigger

Annotate

Annotated : 9 / 91 / 100

↻ Refetch element

Selection

random



On

untagged



Filter

Enter a regex / CONTEXT= for context

► Active model

No model trained

► Advanced options

Très bonne question de @GTabard ? On oscille entre l'irrationnel et l'hypocrisie première langue...

#ReformeRetraites « GTabard: Comment est-on passé de députés @lesRepublicains réclamant la retraite à 65 ans, aux mêmes députés prêts à voter une motion de censure avec @FranceInsoumise et le @RNational\_off ? défenseurs de la retraite à 60 ans ? #coherence #retraites —

<https://twitter.com/gtabard/status/1636441235484475392> »



agressif 1

non-agressif 2

# ActiveTigger

## Validate

To monitor your progress, you have several options.

You can train “quick models” as you go through the annotation process to monitor your progress. This will also help you speed up annotation through **active learning**.

You can then train full BERT models (iteratively) until you are happy with the result.

# ActiveTigger

<http://activetigger2.eschultz.fr/>

Download data and see  
logins in shared Drive folder

[https://emilienschultz.github.io/  
activetigger/docs/quickstart/](https://emilienschultz.github.io/activetigger/docs/quickstart/)

# To go further

## Advanced options & Tricks

1. Active learning
2. Coding Scheme
3. Validation Scores
4. Tips

# ActiveTigger

## Validate

### A quick model

- 1) Select Features under Prepare (sbert)
- 2) Train quick model (using training data) under Tag



Quick model

**10-CV** : 10-fold cross-validation

Evaluates the performance of a model by testing it on 10 different samples.

**Efficient, but to be taken with a grain of salt**

# To go further

## Active learning

Rare cases can take forever to annotate if you pick them out at random

But it is possible to speed up the process




# To go further

## Active learning

a) Look for words / ideas (**regular expression**)

Selection      On      Filter

deterministic    untagged    **Macron**

► Active model  No model trained    ► Advanced options

J'étais présent ce jour avec @FrankGiletti à #Toulon pour la Prise d'armes présidentielle et au discours sur la Revue Nationale Stratégique du Président Emmanuel **Macron**. #Sécurité, stratégie et défense des intérêts nationaux ont été au programme.

<https://twitter.com/Boccaletti83/status/1590384706583154688/photo/1>

# To go further

## Active learning

### b) Active learning

After a quick model is trained, we can look at **least certain predictions**


By classifying these cases specifically, we help the model improve faster

# To go further

## Active learning

### Create a new feature

✓ sbert  
fasttext  
dfm  
regex  
dataset

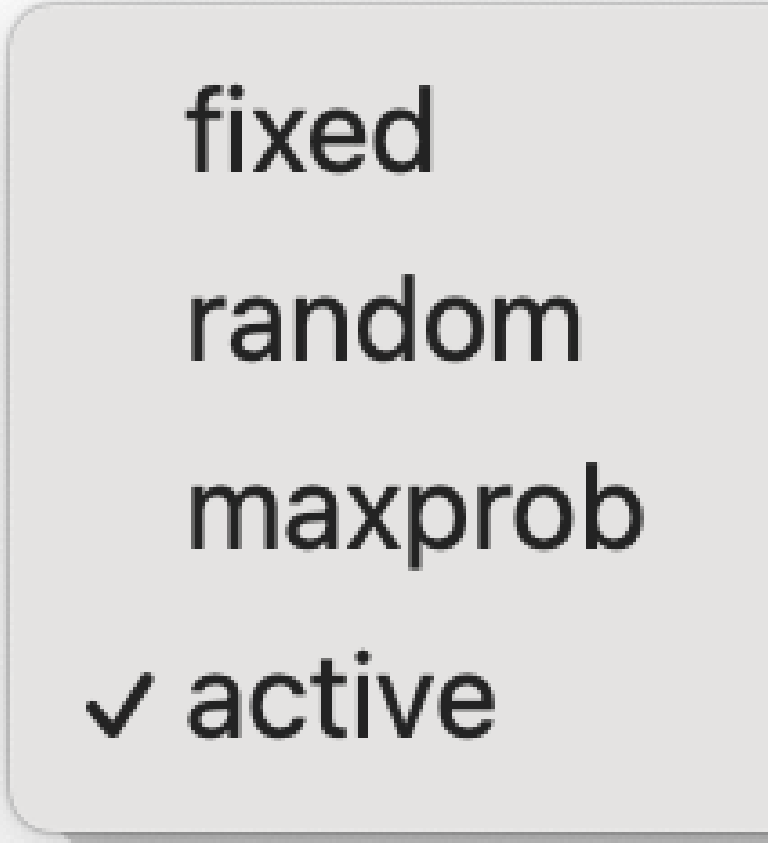
Selection  On  Filter   
► Active model  liblinear Advanced options

@marietouss1 @MaudBregeon Non. La société civile qui entre à l'assemblée avec ses compétences et les met au service du travail parlementaire. Je sais que ça vous dépasse.

Predicted : non-retraites (proba: 0.86)

# To go further

Choose annotation strategy



fixed  
random  
maxprob  
✓ active

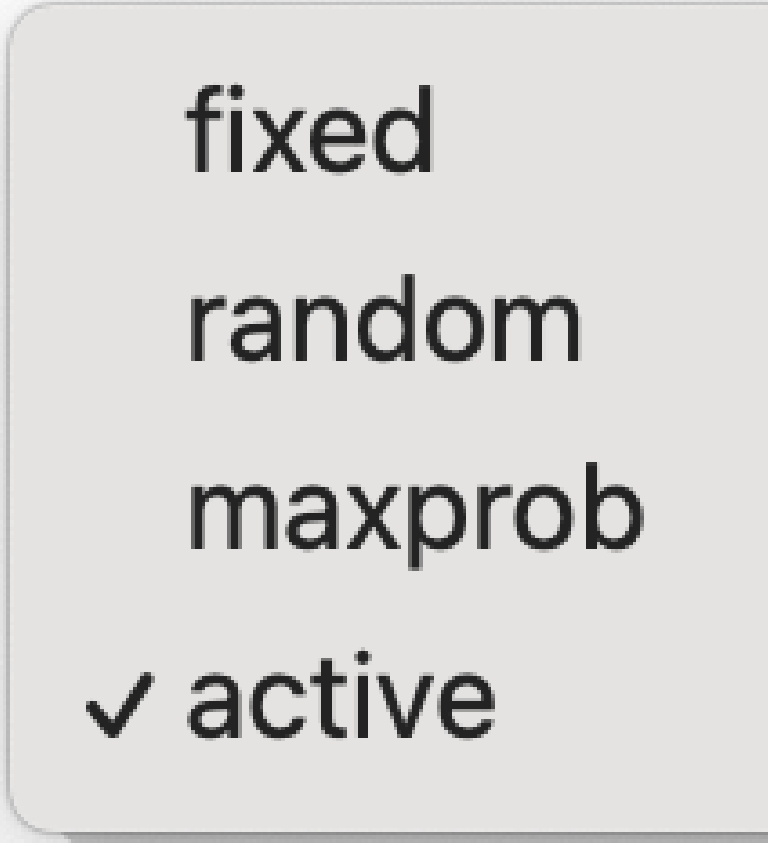
**Fixed:** order of the dataset

**Random:** random draw

→ to see a diversity of cases  
[start with & go back to this often]

# To go further

Choose annotation strategy



fixed  
random  
maxprob  
✓ active

**Active:** most ambiguous phrases  
→ helps to refine model

**MaxProb:** Most certain sentences  
→ helps to consolidate a category

# To go further

## Validation Metrics

▼ Scores	
Key	Value
F1 micro	0.926
F1 macro	0.926
F1 weighted	0.926
F1	{"Pas_Politique":0.928,"Politique":0.925}
Precision	{"Pas_Politique":0.973,"Politique":0.883}
Recall	{"Pas_Politique":0.887,"Politique":0.972}
Accuracy	0.926

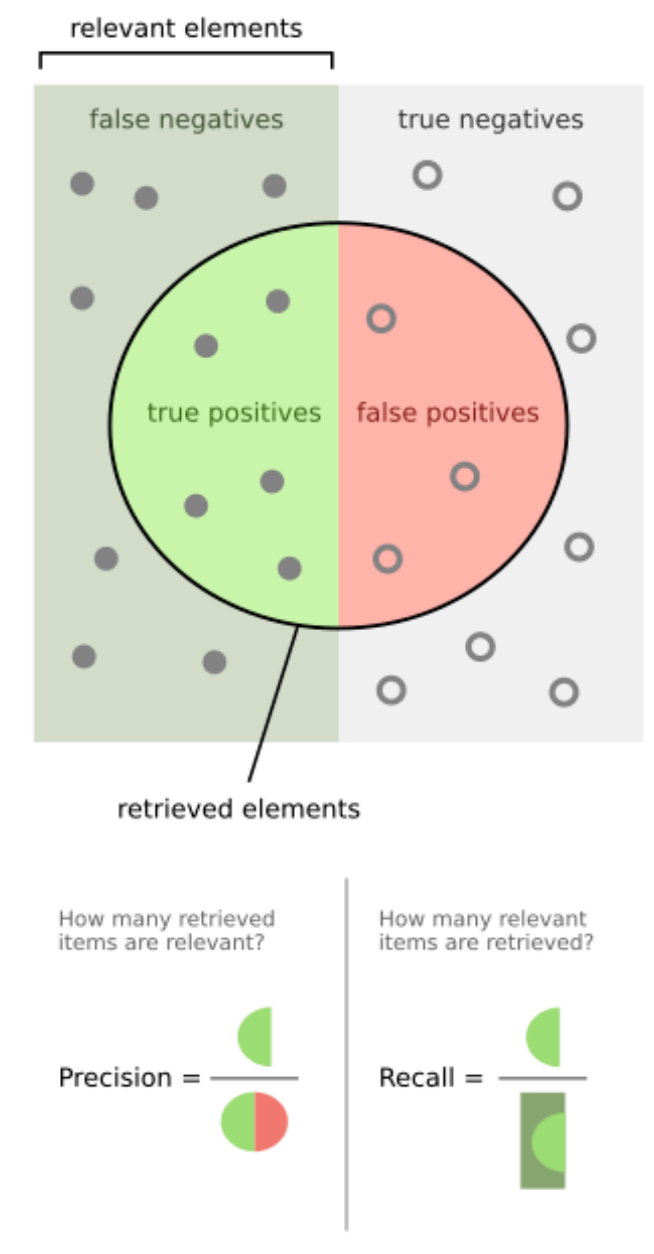
# To go further

## Validation Metrics

**Precision** = Out of all the predicted positives, how many were actually correct?

The model marks 10 emails as spam.  
If 8 of those really are spam, precision is 80%.

👉 Low Precision = Model is a bit trigger happy in labeling cases



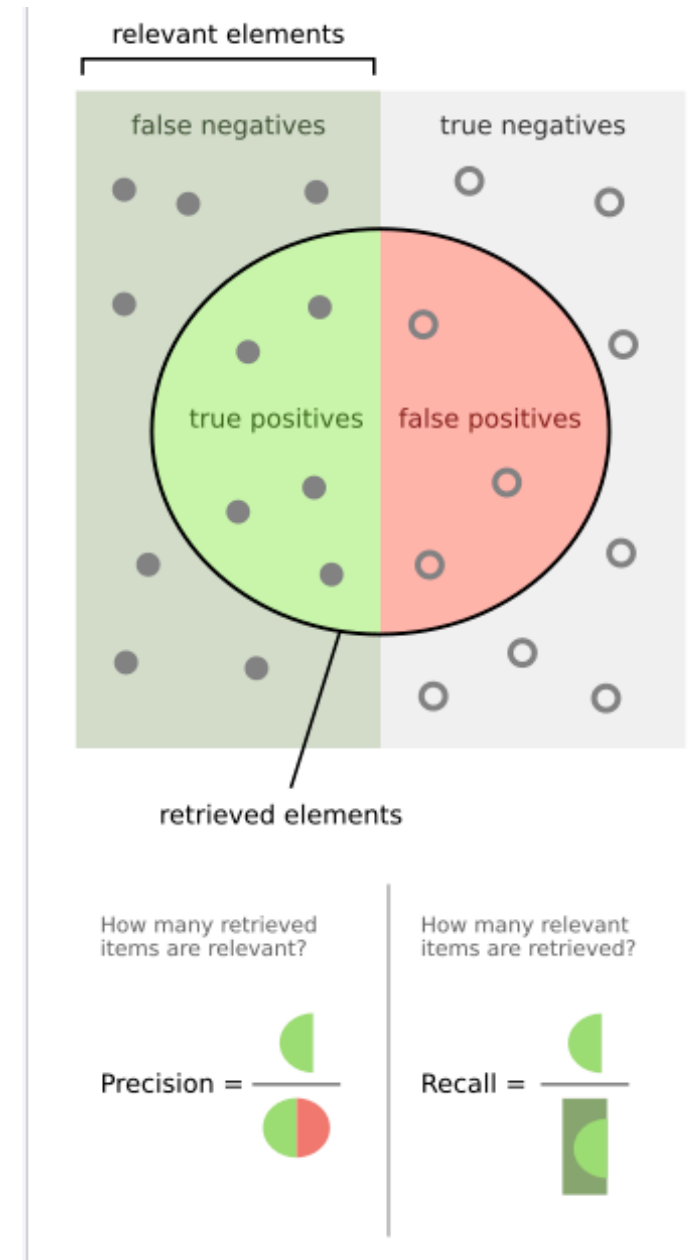
# To go further

## Validation Metrics

**Recall** = Out of all the actual positives, how many did the model correctly identify?

There are 20 actual spam emails.  
The model correctly finds 15. Recall is 75%.

👉 Low Recall means the model misses real positive cases





# To go further

## Validation Metrics

**F1** = Harmonic Mean between Precision and Recall.

No clear threshold, as quality depends on the difficulty of the task (and on how confident you are with the error rate)

# ActiveTigger

## Predict

If you are happy with your test results, you can use the model to predict on the entire corpus, and export them

 Project

 Tag

 Fine-tune

 Export

# ActiveTigger

Export

*Download data (annotations, features, predictions) and fine-tuned models*

Select a format

csv



## Annotations

Export training data

# To go further

## Tips

### How many categories?

- Overall, binary classifiers (“political”/”not political”) are easier to train than multiclass.

### Minimum size for a category?

- **A few %...** otherwise might take some time to train
- A **few hundred annotations** is probably necessary for any classifier to work

# To go further

## Tips

### What if my results are not good?

- Add more data to the training set
- Is your test set separate?
- Does your category make sense? Do you hesitate a lot?
- How do the training data parameters look like?

### How to define a category

- Depends, but
  - Can you explain it easily?
  - Does it take a long time for an expert (you) to annotate it?

# Feedback

We are still working on the interface

Something unclear? Something buggy? A suggestion? Join the beta testers!

We have a Discord for discussion & troubleshooting