



Active Tigger

AN INTUITIVE & COLLABORATIVE
TEXT ANNOTATION TOOL FOR
OPEN SOCIAL SCIENCE RESEARCH



Contact and Information!

Context

Across social sciences, researchers are increasingly interested in skyrocketing amounts of textual data. Even as computer scientists have developed useful tools for analysing textual data, these often place high demand on technical mastery. Widening access to these research methods requires evidence-based, inclusive and transparent tools.

Goals

- Facilitate **collaborative** human annotation and **exploration** of textual datasets
- Use AI to carry out **text classification** and information extraction
- **Democratise** computational social sciences practices for other communities
- Create a lightweight, frugal and **open source** application

Use case

How prevalent is gender in French social sciences?

To answer this question we collected 50,000 journal article abstracts from French publications. Each needed to be annotated: "mentioning gender" (in all its different dimensions) or not.

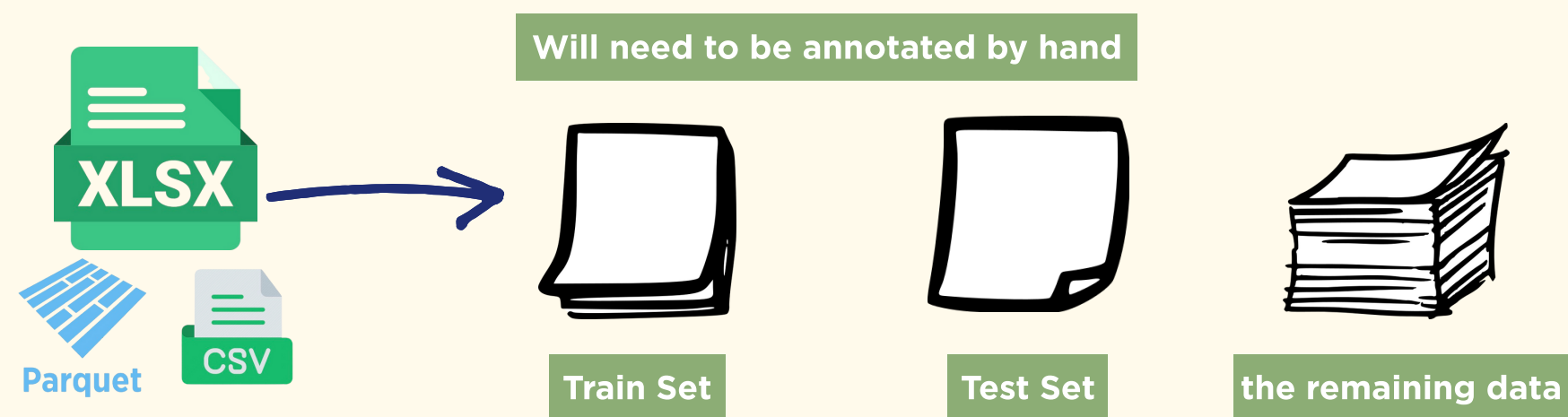
Using Active Tigger helped in many ways:

- The exploration tools eased the stabilisation of a codebook
- The collaborative annotation helped coordinate the process
- After annotating only about 500 abstracts (1% of the dataset), we trained a classifier that performs as well as an expert (F1>0.9)
- We used the model to predict the labels of each abstract

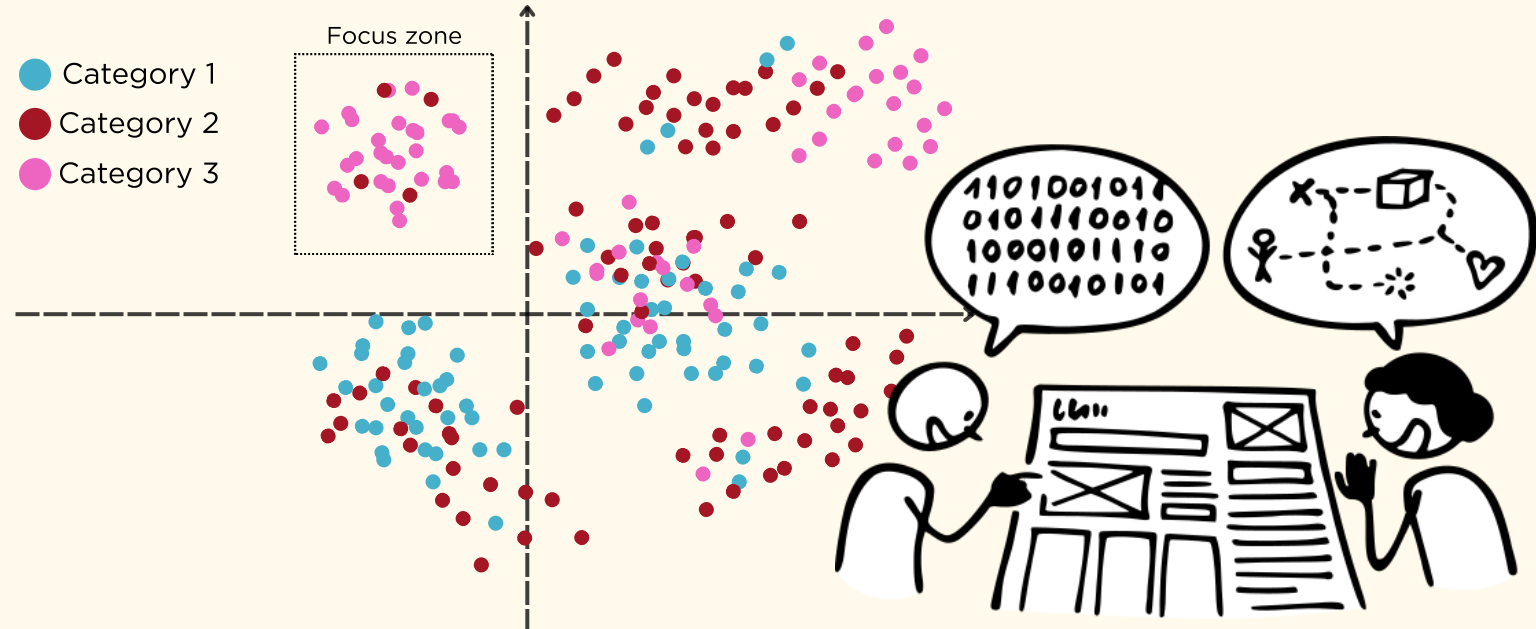
We are actively looking for more use cases. Contact us!

Workflow

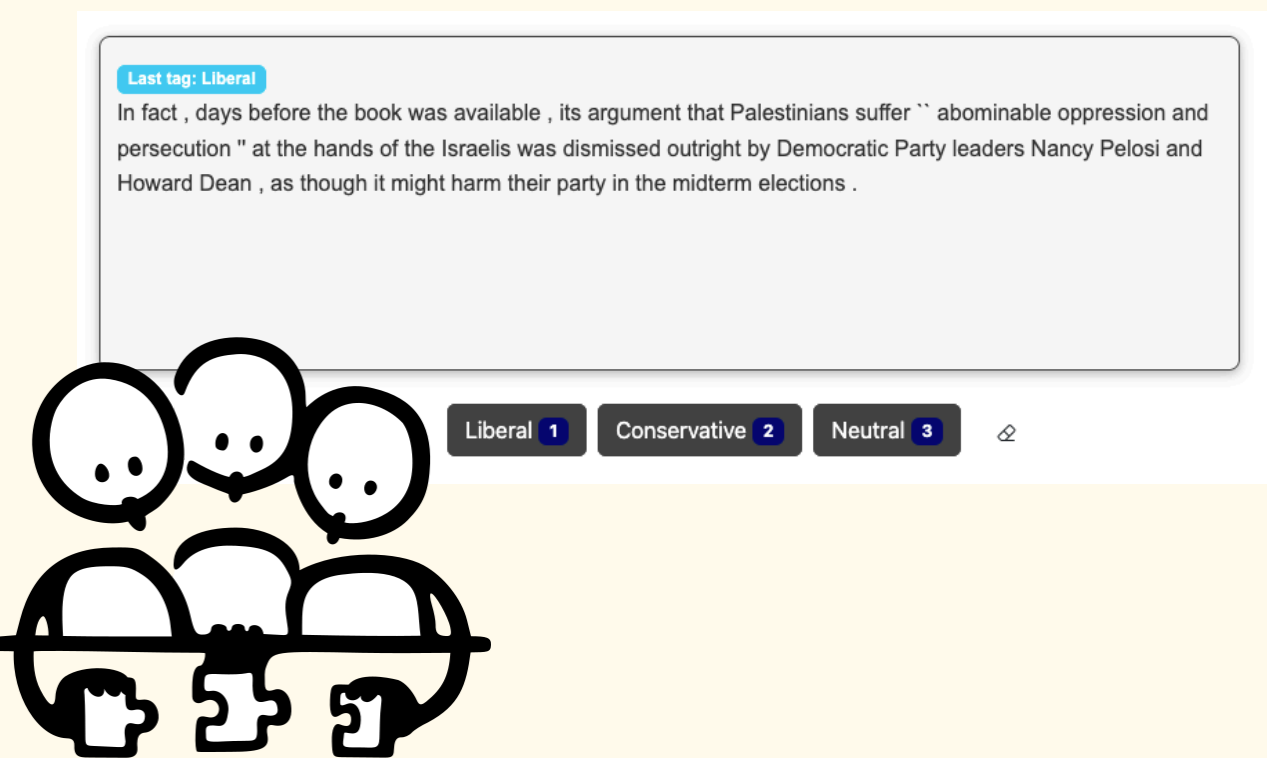
Step 1: Create a project



Step 2: Explore your dataset



Step 3: Annotate your dataset



speed up the learning process with Active Learning

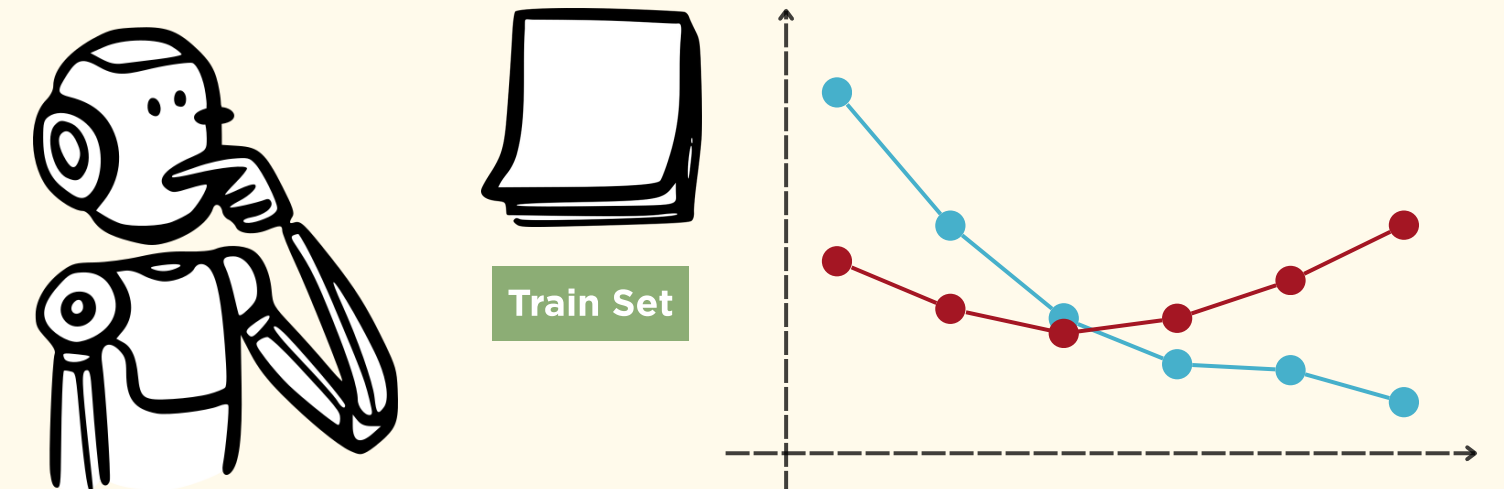
Active Learning is a strategy to choose the best elements to annotate, either by focusing on difficult cases, or on the easiest.

Active Tigger offers other element-selection strategies:

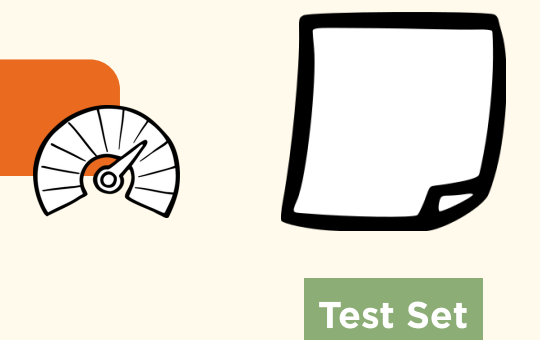
- by filtering with regex patterns
- by selecting an area of the data-visualisation

The Augmented Social Scientist path

Step 4: Train a classifier



Step 5: Test your model



Step 6: Augment yourself



Use your annotated dataset for your research



Use case: we divided the annotation time by 60.
From 340h (estimated) to 6 hours.

How to use:

- From a user-friendly web app or a Python package
- Connect to an existing instance running on a server
- Run own your API (docker and code available in the repository)

How was it made:

- State of the art Python libraries (PyTorch, Hugging Face)
- Built to leverage acceleration of GPUs
- Open source development for transparency and reproducibility

Where to start:

- Read the *Quick Start* and documentation
- Visit the GitHub repository
- Test the app with a testing account

And more to come!

Where we are now
• The first stable version will be released in June 2025

- Next steps
- Generative features
 - Enhanced data exploration with topic analysis
 - Better collaboration management options

You want to contribute? Brilliant!
Please reach out!

Glossary

- Classifier / Classification**
- In machine learning, "classification" refers to a task consisting of labelling texts with predefined labels. In the context of Active Tigger, classifiers are used in the annotation automation.
- BERT models (Bidirectional Encoder Representations from Transformers)**
- Pre-trained models based on the Transformer architecture. These are trained to fill gaps in a sentence on large text datasets.
- Generative LLMs (ex: GPT, Gemini, Llama...)**
- Pre-trained models based on the Transformer architecture. These are trained to predict the next token on gigantic text datasets.
- GPU (Graphics Processing Unit)**
- Computer components designed to accelerate parallel computation. They are broadly used for using and training AI models



Authors: Axel Morin, Émilien Schultz, Annina Claesson, Arnault Chatelain, Emma Bonutti, Julien Boelaert, Étienne Ollion.

ActiveTigger is developed in the CREST laboratory / CSS@IPP with the help of OueWare, and funded by DRARI Île-de-France and Progéo