# Natural Language Processing Final Project

Creation of a Machine Learning System dedicated to Toxicity Detection

## Project Overview

In this final project, you will work in teams to create an almost production-ready machine

learning system for toxicity detection in text, based the Google Jigsaw dataset. While model

training is a core component of this project, you'll also have to take other elements into

consideration, such as scalability of your solution, fairness or explainability.

## Learning Objectives

By completing this project, you will:

• Design and implement a complete ML system for text classification, from model creation

to deployment

• Get a first glimpse of what is called MLOps

• Consider ethical implications in ML deployment

• Work on a complex topic that hasn't been solved yet!

Team Structure

• Teams of 2 members

Project tasks

This project revolves around 2 key aspects:

1. 2. Creation of a machine learning model dedicated to toxicity detection in texts

Exploration of other components required to create a full-fledge ML system. As such, the content of the project consists in the following tasks:

1. 2. 3. Define the metrics you'll use to evaluate the quality of your models

Create a data preprocessing pipeline to prepare the texts for model training

Use a LLM to create a baseline out of your test set

a. You can experiment with multiple LLMs but don't spend too much time on complex prompting experiments: the goal is to have a quick baseline to beat / reach and see how training a custom model compares to it

b. If you don't have a GPU on your own machines such inference will likely take time. You can use Google Collab free GPU notebooks to speed up the inference. 7B / 4B models should run fine if quantized4. 5. 6. 7. c. If that still takes too much time, consider creating a sample out of your testing data.

In that case, use a seed to ensure reproductibility and use the same test sample to compare the performance of your trained model(s) later on Finetune the RoBERTa-base model using the training data attached to this project

a. This step involves hyperparameter optimization, but you don't need to use PBT

optimization approaches seen in class

b. You will need to create a Model Card for your finetuned model

Question :

• Conduct a bias / fairness assessment of your model

o What do we call bias / fairness in ML systems, specifically for texts?

o Measure how much your model is sensible to identified bias

• Estimate how much your baseline / finetuned model is resilient to obfuscation / jailbreak

o Explore how you can mitigate jailbreaking / obfuscation attempts of NLP system

o Implement some approaches in the API that expose your model

o Note: we're talking here of jailbreaking a model, not the API itself. As such,

considerations like authentication / networking security shouldn't be taken into

account.

Deliverables

Your team is expected to submit two main deliverables that will showcase both your technical

implementation and your analytical thinking.

1. Jupyter notebook

• Instructions for running training, inference and evaluation

• Detailed explanation of your project structure

• A model card for your finetuned model

o As models can be quite big, you don't need to send your model The model card will be enough.

2. Technical Report

Your team will prepare a detailed technical report that documents your project. This report

should:

• Be a PDF named PDF named NLP\_Project\_NAME1\_NAME2\_NAME3\_NAME4.pdf,

The said report should include:

• An executive summary highlighting your different results and learnings

• A small section that contextualizes the challenge of toxicity detection in texts

• A walkthrough of your modelling experiments, including:

o Metrics selected to evaluate the quality of your model(s)

o Considerations regarding the LLM baseline creation (model selection,

prompting, results)

o Detailed report on the experiments related to the RoBERTa model finetuning

o Visualization of model performances

* An answer to the questions

Questions / FAQ

If you have any question regarding this project brief, don’t hesitate to contact me at the following

email address: edouard.nadaud@esiea.fr