FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Deep and Reinforcement Learning
2025/2026 - 1st Semester

Emotional Models!

Jorge Henriques
João Correia

1 Introduction

We present here the second practical project, part of the students' evaluation process of the Deep and Reinforcement Machine Learning course of the Master in Informatics Engineering of the University of Coimbra. This work is to be done autonomously by a group of **two students**. The **deadline** for delivering the work is **14th of December of 2025** via Inforestudante. The quality of your work will be judged as a function of the value of the technical work, the written description, and the public discussion. All sources used to perform the work (including the code) must be clearly identified. If you use AI tools in the production of this work (e.g. ChatGPT), you must clearly identify all the parts in which the tool was involved. Please note that, during the defence, you will be required to demonstrate a deep understanding of the content generated by the tool, and this knowledge will be subject to evaluation. The document may be written in Portuguese or in English, using a word processor of your choice[1]. The **written report** is limited to **10 pages long**.

The document should be well structured, including a general **introduction**, **a description of the problem**, the **approach**, the **experimental setup**, an **analysis of the results**, and a **conclusion**. The report should follow the **Springer LNCS format**[2]. The final mark will be given to each member of the group individually. To do the work the student may consult any source he/she wants. Nevertheless, plagiarism will not be allowed and, if detected, it will imply failing the course. While doing the work and when submitting it, you should pay particular attention to the following aspects (whose relative importance depends on the type of work done):

- description of the approach to the problem
- description of the general architecture of the methods used;
- description of the experiment, including a table with the parameters used which should allow full replication;
- description of the evaluation metrics used for the validation: quality of the final result, efficacy, efficiency, diversity, or any other most appropriate;

[1] Latex is preferred

[2] Template available for word and latex under the "Important downloads for authors" at https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines

Do not forget, besides what was just said, that it is fundamental: (1) to do a correct experimental analysis; (2) to do an informed discussion about the results obtained; (3) to put in evidence the advantages of the chosen alternative.

2 Problem Statement

Emotion classification is a central task in Natural Language Processing (NLP). The goal is to determine the predominant emotion conveyed by short text segments such as tweets, social media posts or chat messages. These signals are crucial in a wide range of applications, including opinion mining, brand monitoring, mental health support, conversational agents and human–computer interaction.

Before the advent of large pretrained language models, most emotion classification systems relied on *hand-crafted features*. Typical pipelines used bag-of-words or TF–IDF representations, sometimes enriched with lexicons and simple linguistic features (e.g., part-of-speech tags, polarity cues), combined with classical machine learning models such as *Logistic Regression*, *Support Vector Machines (SVMs)* or shallow *Multi-Layer Perceptrons (MLPs)*. While effective to some extent, these approaches struggled with issues such as vocabulary sparsity, limited generalisation to new domains and a poor ability to capture subtle contextual cues (e.g., sarcasm, negation, multi-word expressions).

The introduction of *Transformer*-based models fundamentally changed how textual information is represented. Models such as BERT and RoBERTa are pretrained on massive corpora and can be reused as *frozen feature extractors* that output dense, contextualised embeddings for each sentence. These embeddings can then be fed into light classifiers for downstream tasks, in a way that is conceptually similar to using a pretrained CNN backbone (e.g., ResNet) in computer vision. Libraries such as `sentence-transformers` further specialise this idea, providing sentence-level embedding models optimised for semantic similarity and downstream classification.

In parallel, the development of *Natural Language Inference* (NLI) models has enabled a different paradigm: *zero-shot classification*. Instead of training a task-specific classifier, a single NLI model can be prompted with the task labels expressed in natural language (e.g., "this text expresses *joy*") and used to perform classification without any additional supervised training. This is especially interesting for emotion detection, where label sets are often small and semantically meaningful.

In this assignment, students will work with the *Emotion* dataset from

HuggingFace, which contains short English texts annotated with six basic emotions: **anger, fear, joy, love, sadness and surprise**. The practical work is designed to mirror a typical image-classification pipeline: first, you will treat pretrained Transformer models as fixed backbones that produce sentence embeddings, training a light classifier on top of these representations. On a second part, you will use an NLI-based model to perform *zero-shot* emotion classification on the same dataset, without any supervised fine-tuning. This parallel setup allows a direct comparison between supervised and zero-shot approaches, highlighting the strengths and limitations of each strategy in the context of modern NLP.

3 Objective

The main objective of this assignment is to introduce students to modern NLP workflows built on top of pretrained Transformer architectures. Through a combination of supervised learning and zero-shot inference, students will gain practical experience with two complementary paradigms currently used in real-world language technologies.

More concretely, by completing this work, students should be able to:

- understand the role of pretrained Transformer models as reusable *feature extractors* for downstream NLP tasks;
- create sentence embeddings using models from the `sentence-transformers` library[3];
- train and evaluate light classification heads (e.g., Logistic Regression or MLPs) on top of frozen Transformer[4] embeddings from last point;
- apply Natural Language Inference (NLI) models to perform *zero-shot classification*, using label descriptions expressed in natural language[5];
- compare supervised and zero-shot performance on the same dataset, identifying strengths, weaknesses and characteristic error patterns of each approach;

[3] https://huggingface.co/sentence-transformers

[4] https://huggingface.co/models?library=sentence-transformers&sort=trending

[5] https://huggingface.co/collections/MoritzLaurer/zeroshot-classifiers

- explore alternative models from the HuggingFace Hub, assessing the impact of model size, architecture and training objective on classification results[6];
- reproduce standard evaluation metrics in text classification, such as accuracy, macro-F1 and confusion matrices;

Overall, the assignment is designed to create a conceptual bridge between traditional machine learning pipelines and modern Transformer-based NLP, while allowing students to acquire hands-on experience with the HuggingFace ecosystem and with two prominent paradigms in contemporary language modelling: supervised learning and zero-shot inference.

Exploring other solutions than the listed ones that are suitable for the problem at hand and considered as extra work, can be as compensation points to cover problems in the listed ones above.

4 Dataset and Evaluation

4.1 Dataset

The dataset used in this assignment is the **Emotion** dataset, available on the HuggingFace Hub under the identifier `dair-ai/emotion`. It consists of short English texts collected primarily from social media, each annotated with one of six basic emotions: **anger, fear, joy, love, sadness, surprise**
The dataset is already split into training, validation and test partitions:

- **Train:** ∼16,000+ examples
- **Validation:** ∼2,000 examples
- **Test:** ∼2,000 examples

These short, informal messages often exhibit ambiguity, emotional nuance and contextual subtlety, making them a relevant benchmark for comparing traditional supervised approaches with modern Transformer-based embeddings and zero-shot LLM methods.

The dataset is loaded programmatically using the following lines of code (check also the project file):

[6] https://huggingface.co/models?pipeline_tag=zero-shot-classification&sort=trending

```
from datasets import load_dataset
ds = load_dataset("dair-ai/emotion")
```

Each entry includes:

- `text`: the input sentence/string,
- `label`: an integer from 0 to 5,
- `label names`: stored in the dataset metadata.

4.2 Evaluation

All models—both supervised classifiers and zero-shot systems must be evaluated on the **same test split** to ensure comparability. Students should report at least the following metrics:

- **Accuracy:** percentage of correctly classified examples.
- **Macro-F1:** average F1 score across all six classes, treating each label equally regardless of class frequency. This metric is especially important for emotion datasets, where some emotions are less frequent.
- **Confusion Matrix:** a 6×6 matrix showing how predictions for each class are distributed across the true labels. This helps identify systematic confusions (e.g., *love* vs. *joy*, or *fear* vs. *sadness*).

Students are encouraged to go beyond the mandatory metrics by analysing:

- qualitative examples of correct and incorrect predictions;
- performance differences across models of different sizes or architectures;
- whether certain emotions (e.g., *love*, *surprise*) are consistently harder for supervised or zero-shot settings;
- potential dataset biases or annotation ambiguities.

All evaluations must be reproducible, and the code used to generate the reported results must be included in the submission.

5 Conclusion

A few short comments. First, the control of the progression of your work will be done during the classes (T and PL). Moreover, you can discuss eventual problems by presenting yourself during office hours. Second, the projects reflect for the most part your current knowledge. The rest will be the object of lecturing soon. Third, we try to balance the difficulty of all the work, but we are aware that this is not an easy task and it is somehow a subjective matter. Fourth, we try to ask for a workload compatible with the value of the work for the final mark.