NLP course 2022

Bonus 1

Text Classification

Prof. Roberto Navigli

Teaching assistants:
Andrea Bacciu, Stefan Bejgu,
Valerio Neri, Riccardo Orlando,
Alessandro Scirè, Simone Tedeschi

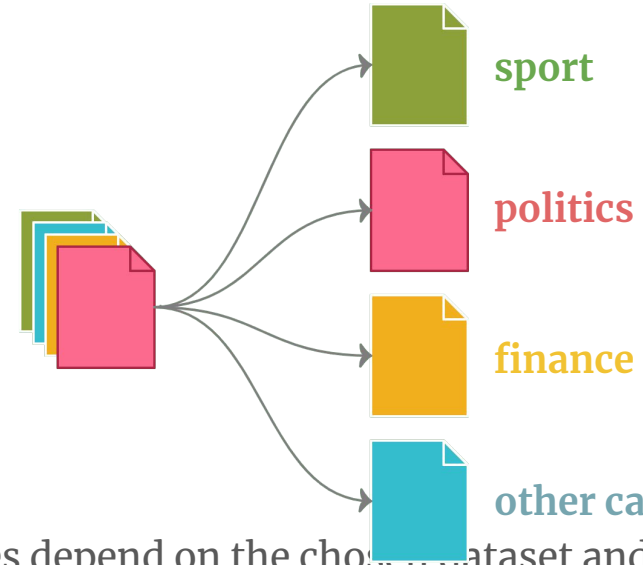
SAPIENZA
UNIVERSITÀ DI ROMA

Text Classification

An introduction

What is Text Classification?

- Text classification is the task of assigning a sentence or document an appropriate category.



- The categories depend on the chosen dataset and can range from topics.

Why is Text Classification important?

- More and more electronic documents become available **every day**!!!
- Such documents represent a **massive amount of information** that is easily accessible
- However, to find information in this huge collection, much work is required to **organize** documents

SAPIENZA NLP

Model

Model: possible approaches

The Text Classification task can be tackled using a wide range of strategies:

- KNN
- SVM
- Word Embeddings (Word2Vec, GloVe, FastText, ...)
- ...

You can choose to implement your own neural architecture or rely on simpler machine learning methods.

Note: for this bonus exercise, **advanced architectures** like LSTMs and Transformer-based architectures **are not allowed**

Dataset and Evaluation

The Dataset

- The dataset is built using news documents extracted from the web
- It contains
  - *186,282* training examples
  - *6844* development examples
  - *6849* testing examples
- The classes are **15**: `business`, `crime`, `culture/arts`, `education`, `entertainment`, `environment`, `food/drink`, `home/living`, `media`, `politics`, `religion`, `sci/tech`, `sports`, `wellness`, `world`
- Example of entry:

```
{
   "text":"The E. W. Brown power plant rides like an ocean liner on a rolling
ridge in Kentucky, its smokestacks and plumes visible across fields…",
   "label":"politics",
   "id":256
}
```

Evaluation

- The evaluation will be conducted on a **BLIND** test set (i.e. you **don't** have the labels for the documents in the test set)
- The **metric** used to evaluate your system is the **error rate**, defined as

$$\text{Error rate} = 100 - \text{accuracy score}$$

- Use the validation split to select the best model/best hyperparameters configuration
- The **final mark** for your submission will depend on the error rate, compared to the error rates achieved by other students
- This bonus exercise will give you up to **1.5 extra points** on top of your final grade

What we expect from you

- Two files **dev.tsv** and **test.tsv** with predictions on the **DEV** and **TEST** sets in a TSV format:
- File example:
  - ID{TAB}predicted_label

```
83900   business
198485  home/living
196171  food/drink
198897  environment
199477  sports
197372  culture/arts
199347  culture/arts
197014  crime
198978  environment
82136   home/living
```

- Your implementation
  - Send **all** the code you use! It should be **reproducible**

Warnings

Things you should be aware of

Please be aware that

This is an **individual exercise!** Collaboration among the students is **not** allowed.

We will check for **plagiarism** both manually and automatically.

It is **not allowed** to:

- Copy from other students
- Share your code with other students
- Copy from online resources (StackOverflow, GitHub, Medium, Kaggle and so on).

However, you are allowed to use material from **external sources** as long as it is **not central** to the homework.

Use of external data

- For your experiments, **use the provided data** (train and dev) in the data folder; use each file as defined in the standard ML conventions (train for training, dev for model selection).
- **Use only the training set to train the model that you submit for evaluation**. If you train it on more data (dev set or any other external data), it will be a **FAIL**.

Tips

A few tips to organize your work:

- Start **as soon as possible**!
  - Training a neural network requires time, possibly hours, depending on your hardware
- Start **small**!
  - If you don't get decent results with a very very simple neural network, there is a good chance that adding other things won't make your model perform better
- Leave some time for **hyperparameter** tuning!
  - Sometimes good hyperparameter combinations can do wonders for your neural network
- Use **Google** Colab (free GPUs!)

Deadline

When to deliver what

Deadline

Submission date: **Mar 20th, 2022 (Sunday)**
23:59:59 Italian time (UTC + 1)

Submit the homework through the submission form on Google Classroom. You have to fill the form with the requested information and a link to the zip folder of the homework on Google Drive.

Late submission is not possible for bonus exercises.

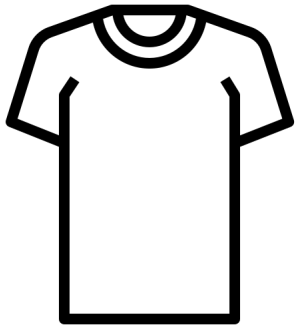
Awards

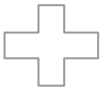
Get a **Sapienza NLP™** t-shirt

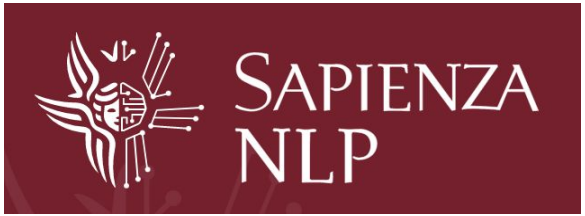
Win a Sapienza NLP t-shirt!

We will hand out amazing Sapienza NLP t-shirts to the **overall top-5** students!

The final ranking will be computed according to the scores on our **secret** test set.







SAPIENZA
NLP

That’s not all

If your work is novel, interesting and original, we will gladly invite you to work together with us to extended on a fully-fledged paper for **TOP-TIER INTERNATIONAL CONFERENCE**!

Just over the last 12 months, the Sapienza NLP group published more than a dozen of papers!

Questions?

If you have a question that may interest your colleagues, **please ask it on Google classroom.**

Otherwise, for personal or other questions, send an email to **ALL** of us (but please, only reach for things that can't be asked on the Google Classroom).

Our emails are:

{bacciu, bejgu, neri, orlando, scire, tedeschi}@diag.uniroma1.it

Good Luck!!