Assignment 2

Identify text written by AI

Teoria Algorítmica da Informação

10/05/2024

Grupo 7

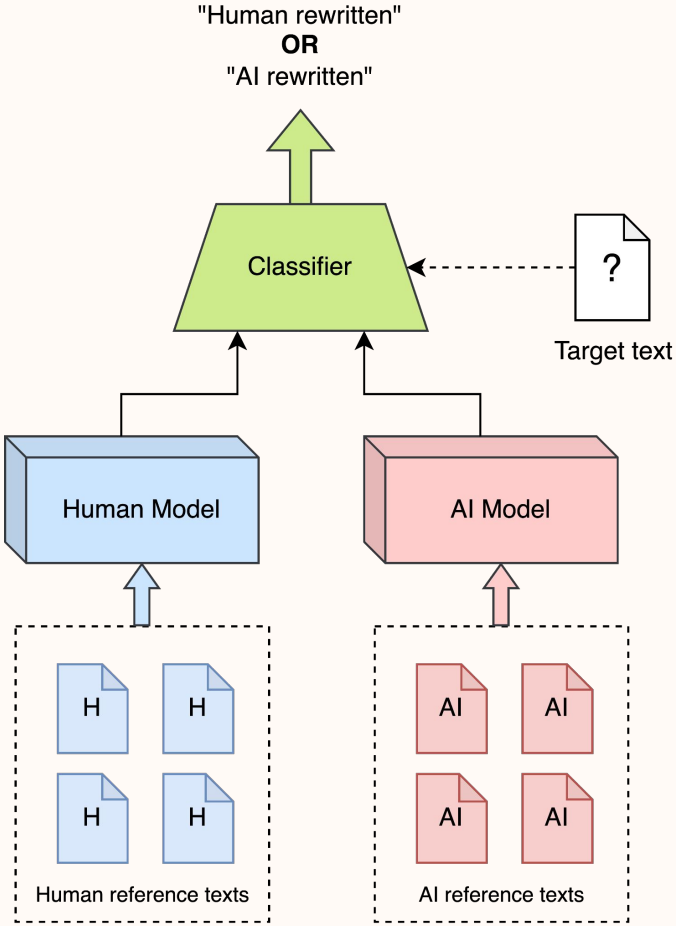
Diogo Magalhães	102470
Leonardo Almeida	102536
Pedro Rodrigues	102778

universidade de aveiro
theoria poiesis praxis

Introduction

- AI-generated text can be misused and negatively impact the society
- There are ML models trained to classify texts as “Human written” or “AI written”
- This models have some drawbacks, for instance high computationally costs

Goal:
Develop a classifier capable of detecting if a text is “Human written” or “AI written” using finite-context models.



Strategy and Implementation

Two main tasks:

- Model creation
- Text classification

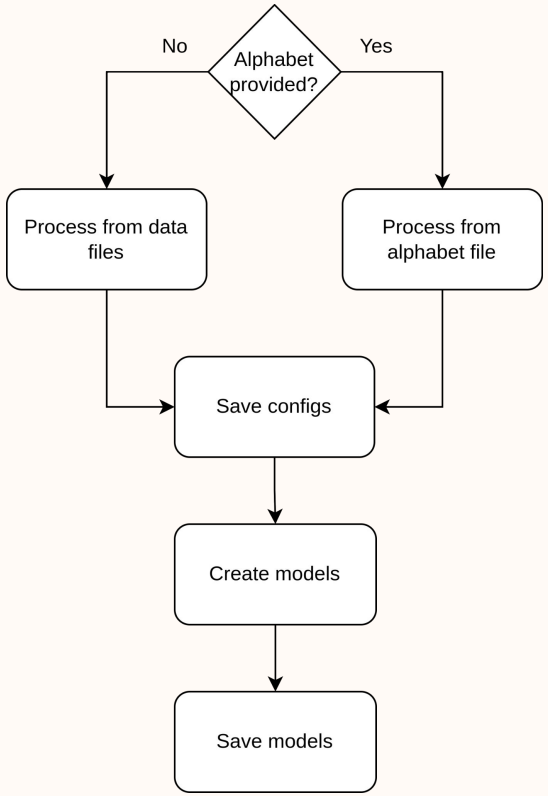
Goal:

The same model can be used multiple times to classify different texts, without the need to retrain it every time.

Model creation

`train.cpp`:

- **-h:** The path to the human reference text file.
- **-g:** The path to the AI-generated text file.
- **-o:** The path to the output folder where the models and configurations will be saved.
- **-a:** The file with the alphabet to be used in the model, if not provided, the program will generate the alphabet based on the input texts.
- **-k:** The order of the Markov model to be used.

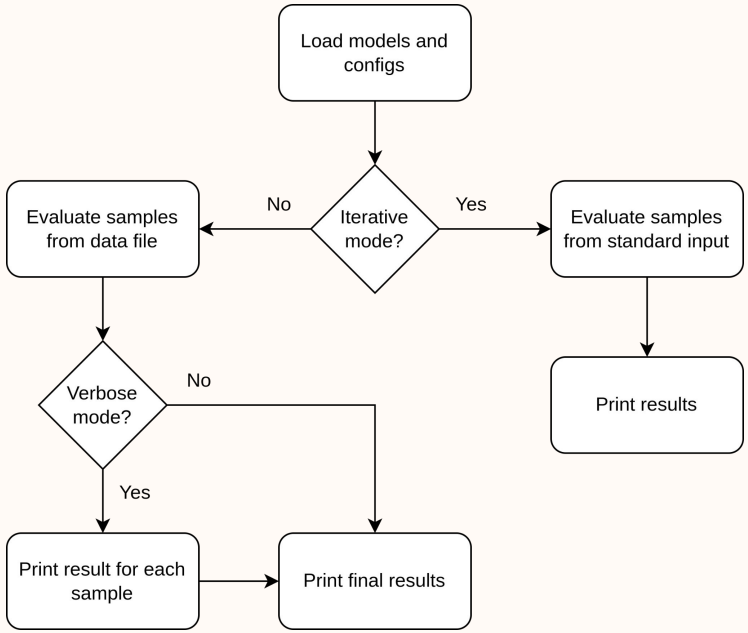


Program workflow

Classification

`was_chatted.cpp`:

- **`-m`:** The path to the models folder.
- **`-d`:** The data file to evaluate (1 sample per line), if not provided, the iterative mode is enabled.
- **`-a`:** The smoothing factor (alpha) value to be used to calculate the probability of a symbol in a context.
- **`-v`:** To enable verbose mode.



Program workflow

Datasets

Two public datasets available on *Hugging Face* were used to generate the reference texts to train and test the models:

- HC3 (Human ChatGPT3)
- AI-human-text

HC3

➔ 24 321 samples
➔ Human reference dataset size: **26.8 MB**
➔ AI reference dataset size: **26.7 MB**

AI-human-text

➔ ≈ 463K samples
➔ Human reference dataset size: **372.7 MB**
➔ AI reference dataset size: **373.6 MB**

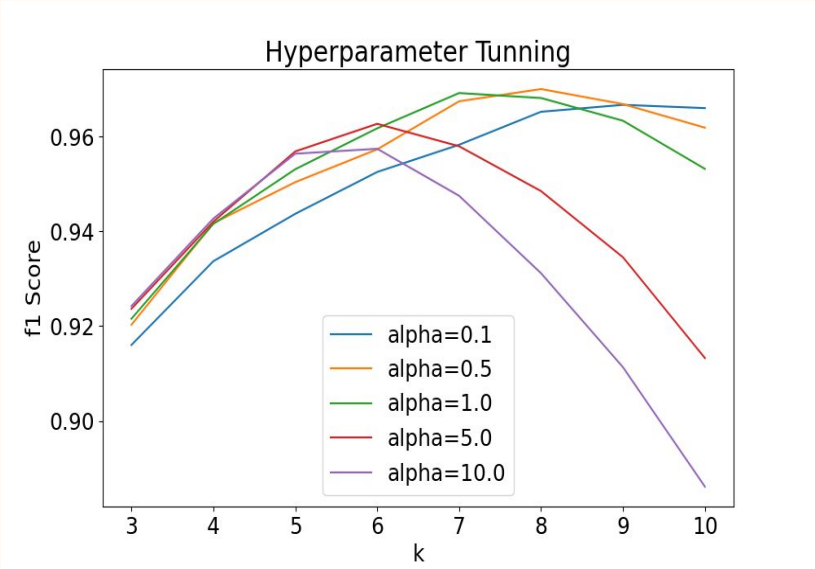
Hugging Face

90% train
10% validation
10% test

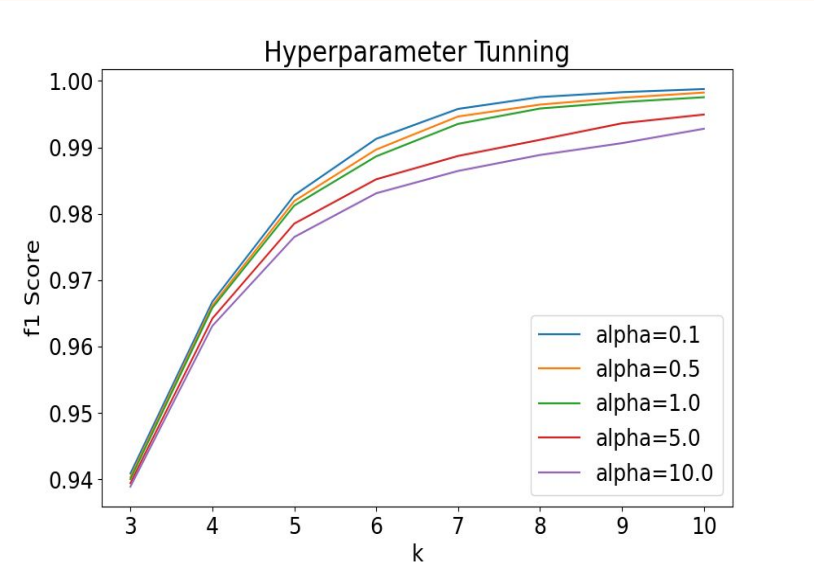
Parameter analysis

We identified 3 tunable hyperparameters: ***k***, ***alpha*** and ***alphabet.***

Our approach was to made a grid-search to find the best values of ***k*** and ***alpha*** and then experiment with different ***alphabets***. We used the following values: ***k*** = [3, 4, 5, 6, 7, 8, 9, 10], ***alpha*** = [0.1, 0.5, 1, 5, 10].



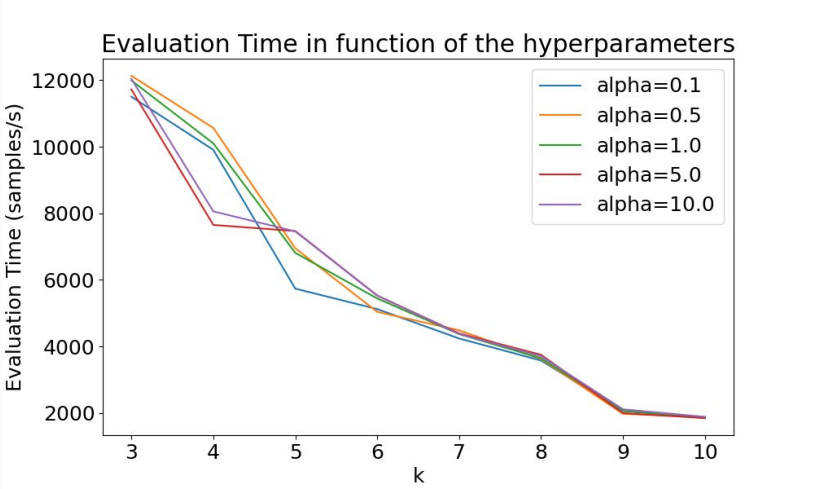
HC3 Dataset



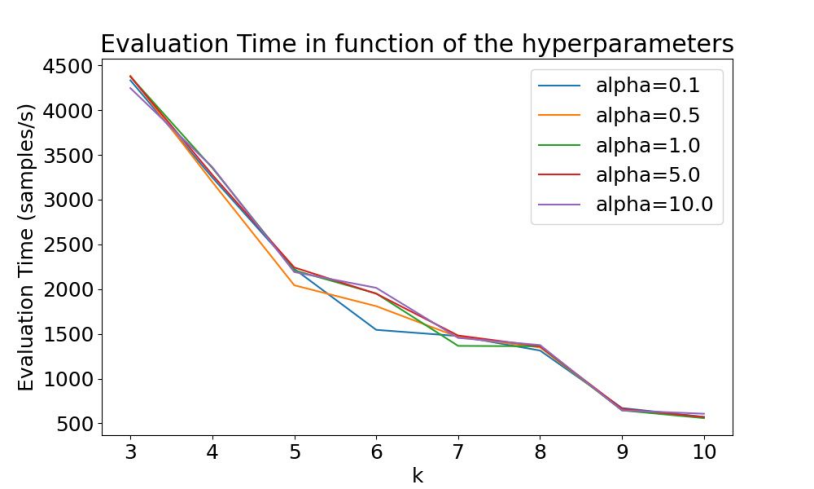
AI-human-text Dataset

Parameter analysis

Not only from hits and misses is a model evaluated, but also from its time performance.



HC3 Dataset

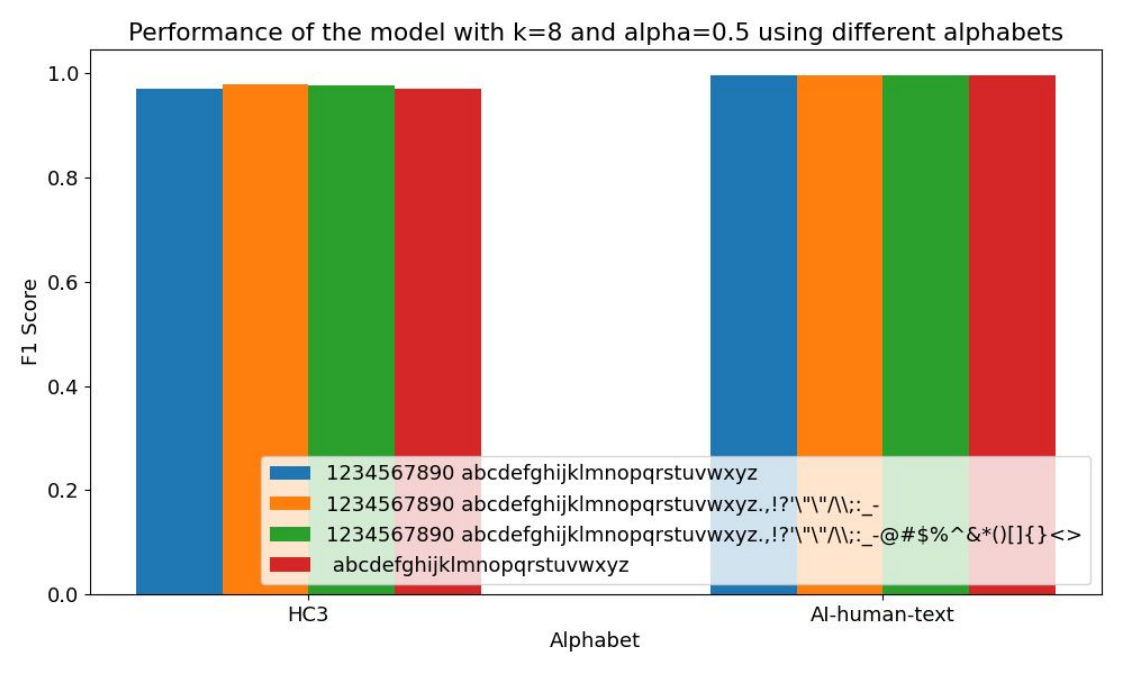


AI-human-text Dataset

We ended up selecting as hyperparameters ***k*** = 8 and ***alpha*** = 0.5 after analyzing both tests.

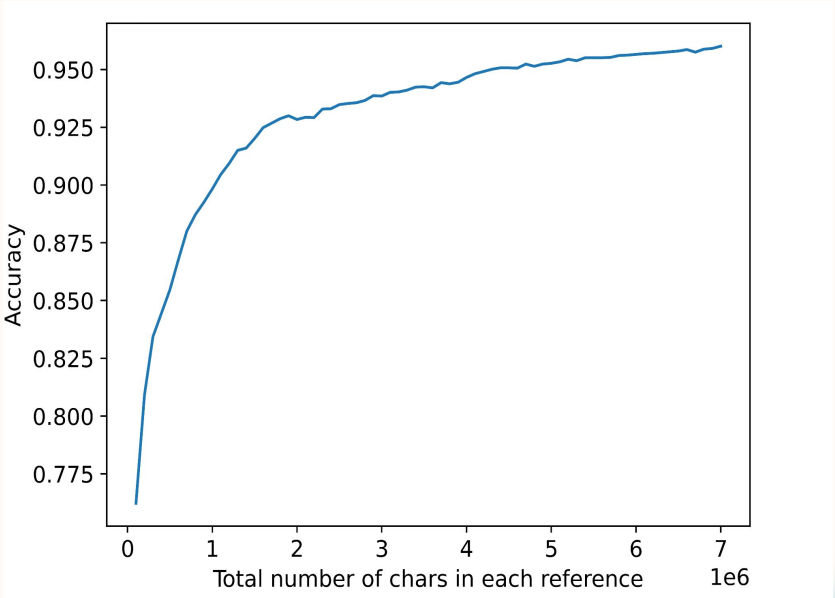
Parameter analysis

After selecting the hyperparameters we tested several alphabets. After analyzing the results we decided to use the alphabet at orange, for both datasets.

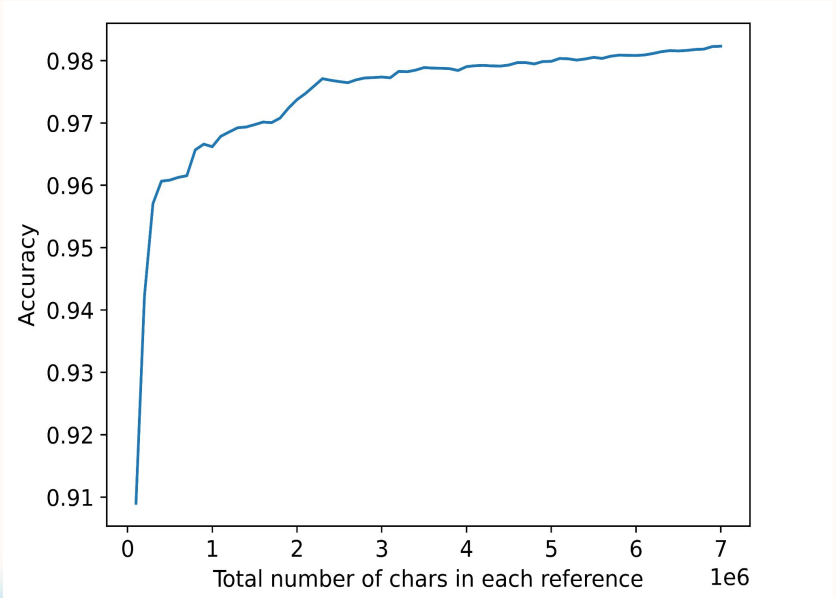


Results - Influence of reference length

HC3 Dataset



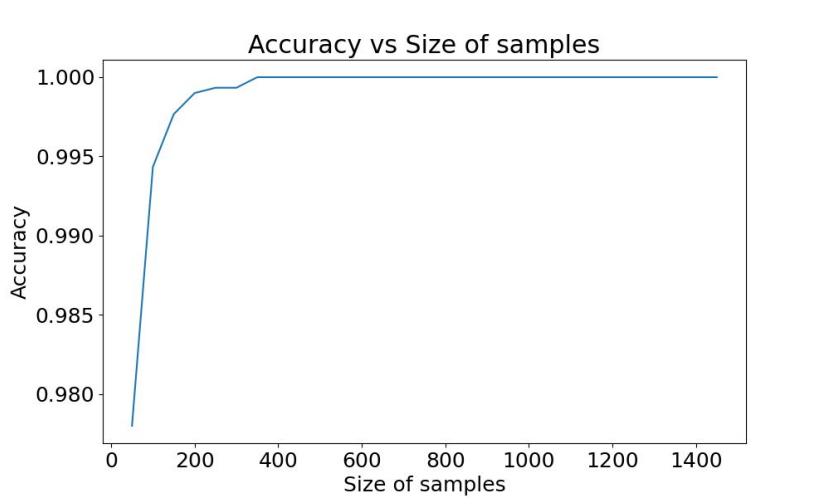
AI-human-text Dataset



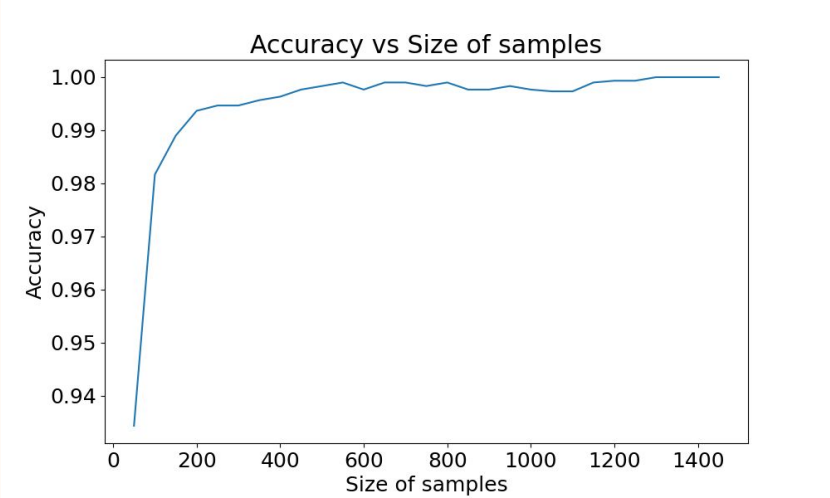
As the length of the references increases, the classifier's performance increases. It tends to increase in a slower pace when reaching 2 millions chars for the first dataset, and 3 millions for the second. **The performance seems to be always increasing at slow rates.**

Results - Influence of target sample

HC3 Dataset



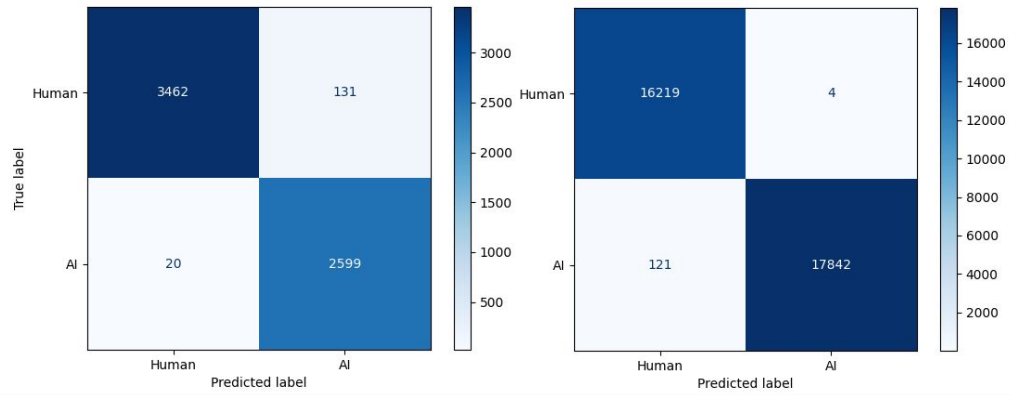
AI-human-text Dataset



The classifier's performance keeps increasing as the number of chars of the target text also increases tending to stabilize after some specific length of target.

Results

Using the test set of the datasets and the generated model we got the following results:



HC3 Dataset **AI-human-text Dataset**

Dataset	Accuracy	F1 Score
HC3	0.97569	0.97521
AI-human-text	0.99634	0.99633

The classifier performed very well and the results outperformed our expectations.

Conclusions

Classifier's performance under various conditions outperformed our expectations.

Reliable alternative to the current transformer based methods.

- Not a black box
- Requires fewer computing power for training and evaluation
- No context (number of tokens) limitation
- Same or better performance