Challenge for "Merit Prize 2025-26"

Portuguese Fake News Detection

Challenge Overview

Build machine learning (ML) models to detect misinformation in Portuguese news articles, and leverage interpretability and explainability methods to analyze results on the FakeNews-PT dataset.

- **Deadline:** 12/12/2025 at 23:59:59 (Lisbon time)
- **Submission guidelines:** Submit a single Gxxx.ZIP in Fenix where xxx is your group number. The ZIP should contain two files: Gxxx_report.pdf with your report and Gxxx_notebook.ipynb with your notebook demo according as for the HWs of the course

Dataset & Resources

The FakeNews-PT Dataset

- **Source:** https://github.com/ro-afonso/fake-news-pt-eu
- **Size:** 60,000+ news articles separated into train / validation / test sets (`.csv`)
- **Task:** Binary classification (0 for fake news, 1 for real)
- **Features:** "Text" (string) and "Label" (integer).

Deliverables

Your submission should be a compressed archive (`.zip`) containing:

1. **Report:** Single PDF document answering all exercises (up to 8 pages). We highly encourage you to write the document in LaTeX and use the NeurIPS official template: https://www.overleaf.com/latex/templates/neurips-2024/tpsbbrdqcmsh (make sure to de-anonymize the document!).

2. **Code:** A repository with your Jupyter notebooks along with a README.md file containing installation instructions and team information. Make sure to add meaningful comments to your code!

References

- A. Wichert, L. Sa-Couto. Machine Learning - A Journey to Deep Learning with Exercises and Answers
- K. Murphy. Probabilistic Machine Learning: An Introduction
- Ch. Bishop. Pattern Recognition and Machine Learning
- M. T. Ribeiro et al., Why Should I Trust You? Explaining the Predictions of Any Classifier. KDD 2016.
- S. M. Lundberg, Su-In Lee: A Unified Approach to Interpreting Model Predictions. NeurIPS 2017.

Academic Integrity

Allowed:

- Standard ML libs (scikit-learn, PyTorch...)
- Course materials, textbooks, papers
- Online documentation
- Team discussions

Prohibited:

- Sharing code with other teams
- Copying without attribution
- Using test labels for training
- Heavy use of AI assistants

Any form of academic dishonesty will result in disqualification.

1 Model Training & Evaluation (10 points)

a) Extract TF-IDF features from the text with a maximum number of features (terms) set to `5000`. Make sure to add smoothing for out-of-vocabulary (OOV) words (`idf smoothing`). Define the minimum and maximum number of documents a term must appear in as `min_df=10`, and the maximum proportion of documents a term can appear in as `max_df=0.9`.

b) Train the following models using 5-fold cross-validation, tune key hyperparameters systematically (e.g., regularization strength λ, tree depth), and document your hyperparameter search process.

- Decision Tree
- Gaussian Naive Bayes
- Logistic Regression with L2 regularization
- Logistic Regression with L1 regularization
- Multi-Layer Perceptron (MLP)[1]

c) Create a comparison table with test metrics: Accuracy, Precision, Recall, and F1-score. For the best classifier, draw its ROC curve and compute AUC.

2 Model Interpretation (7 points)

a) For your best Logistic Regression model, extract and visualize the weights in a bar plot:

- Top 10 words most indicative of fake news
- Top 10 words most indicative of real news

b) Compare L1 vs L2 regularized models: How many features have non-zero weights in each? What does this tell you about feature selection? When would you prefer L1 vs L2 regularization for text classification?

c) For your best Logistic Regression model, select samples in the validation set with ID `2921, 2437, 5557, 1697`, and extract explanations with LIME (Ribeiro et al., 2016; Lundberg and Lee, 2017).

- For a practical reference, check this tutorial.

d) For your MLP, select samples in the validation set with ID `2921, 2437, 5557, 1697`, and extract explanations with LIME and permutation importance. For permutation importance, select 1K random samples. Visualize the results and discuss their differences.

3 Clustering (3 points)

a) Apply K-Means with K=5 on your training set.

b) Inspect 3 documents closest to each centroid, and afterwards, assign semantic labels to each cluster (e.g., "political fake news", "health misinformation").

c) Visualize clusters in 2D using PCA. Create two plots: one colored by cluster assignment, one by true label.

[1] You are free to choose your architecture but up to 2 hidden layers.