Noun Gender Experiment Workflow

Start

⬇

Step 1: Select Models for Comparison

Use **same 10 models as Ekaterina**

FlauBERT (small cased, base uncased, base cased, large cased)

CamemBERT (base cased)

XLM-R (base cased, large cased)

mBERT (base cased, base uncased)

DistilBERT (base cased)

⬇

Step 2: Extract Word Embeddings

Generate **word embeddings** for **French nouns** from each model

⬇

Step 3: Train Perceptron Classifier (Baseline)

Train a **simple perceptron** classifier

Train-test split: 80% training, 20% testing

Evaluate baseline accuracy

⬇

Step 4: Apply SHAP/LIME for Feature Selection

Run SHAP and **LIME** separately to rank embeddings by importance

Sort features by impact scores

⬇

Step 5: Select Top-N Important Features

Pick **N impactful embeddings** (e.g., **top 20 or 30**)

⬇

Step 6: Retrain Perceptron with Selected Features

Train perceptron using **only selected embeddings**

Evaluate new accuracy

⬇

Step 7: Compare Results

Compare accuracy of:

Baseline (all embeddings)

SHAP/LIME-selected embeddings

Compare impactful embeddings with **Ekaterina's results**

⬇

Step 8: Analyze Performance & Insights

Check **accuracy improvements/losses**

Identify **shared vs unique important embeddings** with Ekaterina's results

⬇

End

Reference Materials for Comparison (Noun Gender Experiment)

- Ekaterina's Experiment Code: GitHub Link
- Ekaterina's Accuracy Results: GitHub Link
- Ekaterina's High-Impact Dimensions: GitHub Link