Classificació de textos

Millores per a la Xarxa Neuronal de Classificació de Text

1. Millora de la Representació del Text (Embeddings)

- **Embeddings Generalistes i Específics:** Per millorar la precisió, pots usar embeddings com Word2Vec, GloVe o FastText, que capturen significat semàntic. En classificació de sentiments, embeddings preentrenats especialitzats com BERTweet poden captar millor les emocions.
- **Fine-tuning de Transformers (BERT, RoBERTa):** Els transformers com BERT, o versions com DistilBERT (menys exigent computacionalment), permeten afinar-se per tasques específiques, com la detecció d'idioma o la classificació de sentiments. Aquests models capten bé els matisos del llenguatge, ajudant a identificar sentiments i característiques d'idioma.

2. Augmentació i Balanceig de Dades

- **Augmentació de Textos:** Generar noves frases amb sinònims o canviant l'ordre d'algunes paraules pot ajudar a enriquir les dades per millorar la robustesa. Aquesta tècnica és especialment útil per a la classificació de sentiments, on es poden crear variants positives, neutres i negatives del mateix text.
- **Balanceig de Classes:** Mantenir un equilibri entre classes (`eng/other` per a idioma; `positive/neutral/negative` per a sentiments) ajuda a evitar biaixos en el model. Si hi ha menys exemples en una classe, es pot augmentar duplicant o creant variacions d'aquests exemples.

3. Arquitectura del Model

- **Model Seqüencial (RNN, LSTM, GRU):** En casos on els transformers no es poden utilitzar per limitacions de recursos, xarxes com LSTM o GRU poden ser útils per captar la seqüència de paraules en classificació d'idiomes o sentiments.
- **Transformers amb Atenció (BERT, RoBERTa):** Els transformers amb atenció són efectius tant per detectar l'idioma com per classificar sentiments. La seva capacitat per captar context en llargues seqüències de text els fa ideals per a la classificació complexa de textos.

4. Optimització del Model i Hiperparàmetres

- **Optimitzadors i Learning Rate:** Experimentar amb optimitzadors com Adam o RMSprop i ajustar el *learning rate* pot ajudar a millorar la convergència del model. També pots utilitzar un scheduler per ajustar el learning rate durant l'entrenament, especialment útil en xarxes grans com transformers.
- **Batch Size i Early Stopping:** Prova amb diferents *batch sizes* i aplica tècniques d'early stopping per evitar el sobreentrenament, important en tasques amb menys dades.

5. Regularització

- **Dropout i Regularització L2:** El *dropout* ajuda a evitar el sobrefitament, especialment quan s'utilitzen arquitectures complexes com transformers o xarxes multicapa.
- **Data Augmentation amb Soroll:** Afegeix soroll o lleugers canvis en el text per ajudar a millorar la robustesa del model, útil per ambdues tasques.

6. Preprocessament del Text

- **Neteja i Normalització del Text:** Eliminar caràcters especials, convertir a minúscules i eliminar espais en blanc ajuda a millorar la coherència. Aquest pas és important en la detecció d'idioma i sentiments per evitar soroll en les dades.
- **Eliminació de Stop Words:** Depenent de la tasca, eliminar paraules sense significat pot reduir soroll, però en alguns casos és millor mantenir-les per captar l'estructura completa de la frase.

7. Avaluació i Validació

- **K-Fold Cross-Validation:** Utilitzar validació creuada ajuda a comprovar la generalització del model en ambdues tasques. Aquesta tècnica divideix les dades en múltiples grups per assegurar que el model es prova en diferents subconjunts.
- **Mètriques Específiques:** A més de l'exactitud, per a la classificació de sentiments, utilitzar precisió (precision), sensibilitat (recall) i F1 score pot donar una millor visió de com el model classifica cada sentiment.

8. Altres Tècniques Avançades

- **Models Transformers Específics (BERTweet per sentiments):** En la classificació de sentiments, models dissenyats per a text en xarxes socials, com BERTweet, poden millorar significativament el rendiment.
- **Explainability Tools:** Utilitzar eines com LIME o SHAP per entendre les decisions del model pot ajudar a refinar el model segons la seva capacitat de classificar correctament idiomes i sentiments.

Exercici 0

Entrenament, tot bé

```
Evaluating: 100.00% |██████████| 3/3 [00:00<00:00, 111.41it/s]
Epoch 13/20 (Train: loss: 0.29, accuracy: 0.99) (Eval: loss: 0.28, accuracy: 0.99)

Training:   100.00% |██████████| 10/10 [00:00<00:00, 42.67it/s]
Evaluating: 100.00% |██████████| 3/3 [00:00<00:00, 115.69it/s]
Epoch 14/20 (Train: loss: 0.24, accuracy: 0.99) (Eval: loss: 0.23, accuracy: 0.99)

Training:   100.00% |██████████| 10/10 [00:00<00:00, 40.92it/s]
Evaluating: 100.00% |██████████| 3/3 [00:00<00:00, 100.27it/s]
Epoch 15/20 (Train: loss: 0.19, accuracy: 1.00) (Eval: loss: 0.18, accuracy: 0.99)

Training:   100.00% |██████████| 10/10 [00:00<00:00, 42.85it/s]
Evaluating: 100.00% |██████████| 3/3 [00:00<00:00, 115.69it/s]
Epoch 16/20 (Train: loss: 0.15, accuracy: 1.00) (Eval: loss: 0.15, accuracy: 0.99)

Training:   100.00% |██████████| 10/10 [00:00<00:00, 43.17it/s]
Evaluating: 100.00% |██████████| 3/3 [00:00<00:00, 107.43it/s]
Epoch 17/20 (Train: loss: 0.12, accuracy: 1.00) (Eval: loss: 0.12, accuracy: 0.99)

Training:   100.00% |██████████| 10/10 [00:00<00:00, 41.78it/s]
Evaluating: 100.00% |██████████| 3/3 [00:00<00:00, 115.69it/s]
Epoch 18/20 (Train: loss: 0.10, accuracy: 1.00) (Eval: loss: 0.10, accuracy: 0.99)

Training:   100.00% |██████████| 10/10 [00:00<00:00, 43.22it/s]
Evaluating: 100.00% |██████████| 3/3 [00:00<00:00, 107.43it/s]
Epoch 19/20 (Train: loss: 0.08, accuracy: 1.00) (Eval: loss: 0.09, accuracy: 0.99)

Training:   100.00% |██████████| 10/10 [00:00<00:00, 43.59it/s]
Evaluating: 100.00% |██████████| 3/3 [00:00<00:00, 111.41it/s]
Epoch 20/20 (Train: loss: 0.06, accuracy: 1.00) (Eval: loss: 0.08, accuracy: 0.99)
PS C:\Users\Denis\Documents\DAM-Algorismes>
```

Classificació, tot bé

```
105     accuracy = correct / total

PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

Text: Puoi consigliare un buon ristorante?          ..., Prediction: 52.90% = other  (other  > correct)
Text: The sink is clogged                           ..., Prediction: 54.85% = eng    (eng    > correct)
Text: The window won't close                        ..., Prediction: 51.16% = eng    (eng    > correct)
Text: I need to send an email                       ..., Prediction: 55.57% = eng    (eng    > correct)
Text: I need to find an ATM                         ..., Prediction: 55.46% = eng    (eng    > correct)
Text: El concert va ser increïble                   ..., Prediction: 52.87% = other  (other  > correct)
Text: Les escaleres són relliscoses                 ..., Prediction: 52.02% = other  (other  > correct)
Text: L'examen será la semana próxima               ..., Prediction: 55.37% = other  (other  > correct)
Text: I lost my ticket                              ..., Prediction: 54.30% = eng    (eng    > correct)
Text: Can you pass me the menu?                     ..., Prediction: 52.56% = eng    (eng    > correct)
Text: I lost my passport                            ..., Prediction: 53.55% = eng    (eng    > correct)
Text: The shower is broken                          ..., Prediction: 54.07% = eng    (eng    > correct)
Text: The museum is closed today                    ..., Prediction: 55.46% = eng    (eng    > correct)
Text: The car needs to be repaired                  ..., Prediction: 52.77% = eng    (eng    > correct)
Text: La playa está muy llena hoy                   ..., Prediction: 55.55% = other  (other  > correct)
Text: The website is down                           ..., Prediction: 53.70% = eng    (eng    > correct)
Text: Please pass me the salt                       ..., Prediction: 51.93% = eng    (eng    > correct)
Text: Ho bisogno di caricare il telefono            ..., Prediction: 53.27% = other  (other  > correct)
Text: I love reading books in the park              ..., Prediction: 55.46% = eng    (eng    > correct)
Text: The movie starts in 10 minutes                ..., Prediction: 53.10% = eng    (eng    > correct)
Text: La porta no s'obre                            ..., Prediction: 55.66% = other  (other  > correct)
Text: He dimenticat el paraigua                     ..., Prediction: 50.82% = other  (other  > correct)
Text: I have a toothache                            ..., Prediction: 52.85% = eng    (eng    > correct)
Text: Can you give me directions?                   ..., Prediction: 51.57% = eng    (eng    > correct)
Text: ¿Puedes sacarnos una foto?                    ..., Prediction: 51.23% = other  (other  > correct)
Text: ¿Me puedes ayudar a traducir esto?            ..., Prediction: 55.33% = other  (other  > correct)
Text: The food was delicious                        ..., Prediction: 53.00% = eng    (eng    > correct)
Text: Quel est ton numéro de téléphone?             ..., Prediction: 52.81% = other  (other  > correct)
Text: La habitación es demasiado pequeña            ..., Prediction: 55.62% = other  (other  > correct)
Text: ¿A qué hora es el check-out?                  ..., Prediction: 56.45% = other  (other  > correct)

Validation of 50 examples: success: 48/50, accuracy: 96.00, Error rate: 0.04
PS C:\Users\Denis\Documents\DAM-Algorismes>
```

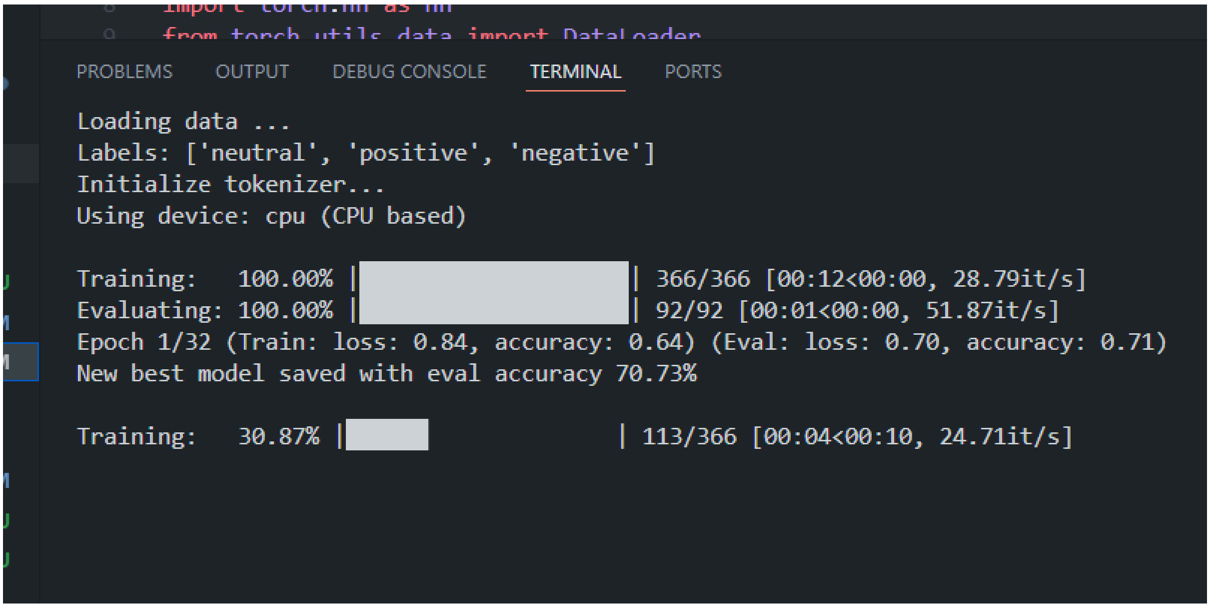
Exercici 1

Entrenament

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

Loading data ...
Labels: ['neutral', 'positive', 'negative']
Initialize tokenizer...
Using device: cpu (CPU based)

Training:   39.34% |████              | 144/366 [00:04<00:08, 26.24it/s]
```



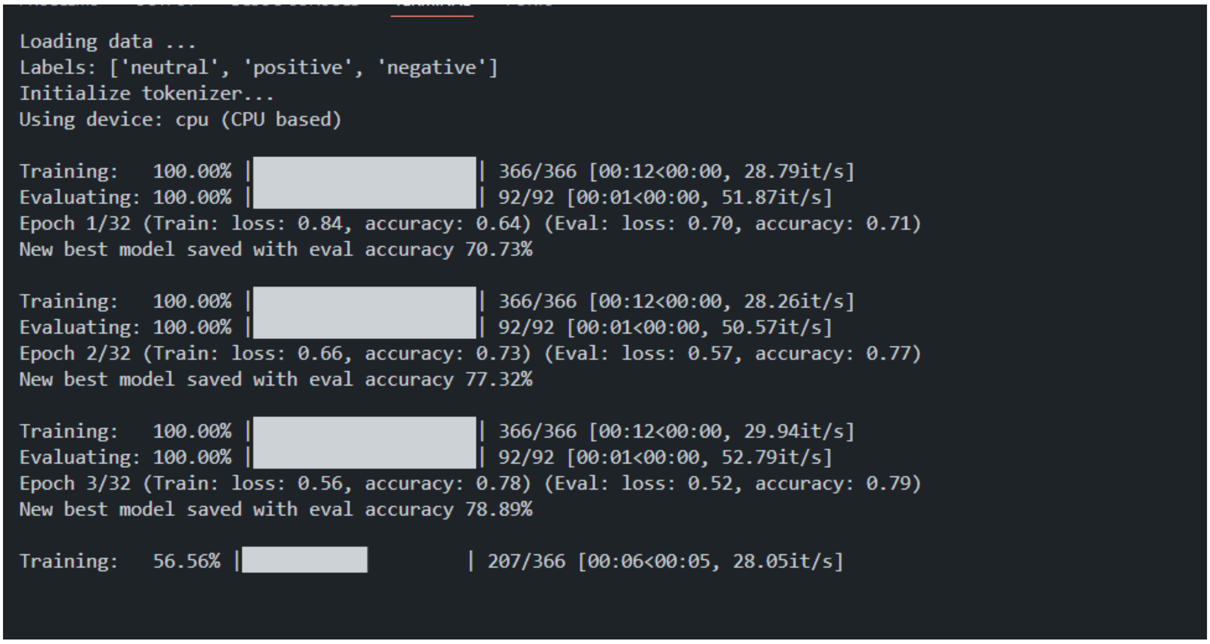
```
import torch.nn as nn
from torch.utils.data import DataLoader
```

PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

```
Loading data ...
Labels: ['neutral', 'positive', 'negative']
Initialize tokenizer...
Using device: cpu (CPU based)

Training:   100.00% |██████████| 366/366 [00:12<00:00, 28.79it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 51.87it/s]
Epoch 1/32 (Train: loss: 0.84, accuracy: 0.64) (Eval: loss: 0.70, accuracy: 0.71)
New best model saved with eval accuracy 70.73%

Training:   30.87% |███       | 113/366 [00:04<00:10, 24.71it/s]
```



```
Loading data ...
Labels: ['neutral', 'positive', 'negative']
Initialize tokenizer...
Using device: cpu (CPU based)

Training:   100.00% |██████████| 366/366 [00:12<00:00, 28.79it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 51.87it/s]
Epoch 1/32 (Train: loss: 0.84, accuracy: 0.64) (Eval: loss: 0.70, accuracy: 0.71)
New best model saved with eval accuracy 70.73%

Training:   100.00% |██████████| 366/366 [00:12<00:00, 28.26it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 50.57it/s]
Epoch 2/32 (Train: loss: 0.66, accuracy: 0.73) (Eval: loss: 0.57, accuracy: 0.77)
New best model saved with eval accuracy 77.32%

Training:   100.00% |██████████| 366/366 [00:12<00:00, 29.94it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 52.79it/s]
Epoch 3/32 (Train: loss: 0.56, accuracy: 0.78) (Eval: loss: 0.52, accuracy: 0.79)
New best model saved with eval accuracy 78.89%

Training:   56.56% |█████▋    | 207/366 [00:06<00:05, 28.05it/s]
```

```
EarlyStopping counter: 2 out of 8

Training:   100.00% |██████████| 366/366 [00:11<00:00, 31.50it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 53.90it/s]
Epoch 10/32 (Train: loss: 0.33, accuracy: 0.88) (Eval: loss: 0.51, accuracy: 0.80)
EarlyStopping counter: 3 out of 8

Training:   100.00% |██████████| 366/366 [00:11<00:00, 31.26it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 48.73it/s]
Epoch 11/32 (Train: loss: 0.31, accuracy: 0.88) (Eval: loss: 0.52, accuracy: 0.80)
EarlyStopping counter: 4 out of 8

Training:   100.00% |██████████| 366/366 [00:12<00:00, 28.36it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 54.39it/s]
Epoch 12/32 (Train: loss: 0.30, accuracy: 0.89) (Eval: loss: 0.53, accuracy: 0.80)
EarlyStopping counter: 5 out of 8

Training:   100.00% |██████████| 366/366 [00:11<00:00, 32.31it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 54.20it/s]
Epoch 13/32 (Train: loss: 0.27, accuracy: 0.90) (Eval: loss: 0.54, accuracy: 0.80)
EarlyStopping counter: 6 out of 8

Training:   100.00% |██████████| 366/366 [00:11<00:00, 32.39it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 52.77it/s]
Epoch 14/32 (Train: loss: 0.26, accuracy: 0.90) (Eval: loss: 0.56, accuracy: 0.79)
EarlyStopping counter: 7 out of 8

Training:   100.00% |██████████| 366/366 [00:11<00:00, 32.43it/s]
Evaluating: 100.00% |██████████| 92/92 [00:01<00:00, 53.45it/s]
Epoch 15/32 (Train: loss: 0.25, accuracy: 0.91) (Eval: loss: 0.57, accuracy: 0.79)
EarlyStopping counter: 8 out of 8
Early stopping triggered
PS C:\Users\Denis\Documents\DAM-Algorismes>
```

Classificació

```
Text: @JetBlue finally! Finally!                        ..., Prediction:  82.00% = "positive" ("neutral" > wrong)
Text: @JetBlue by any chance do u offer fresh guacamole ..., Prediction:  80.73% = "neutral" ("neutral" > correct)
Text: @SouthwestAir this is not a fair set up. I payed f..., Prediction:  99.31% = "negative" ("negative" > correct)
Text: @united every time I search a flight your site log..., Prediction:  99.63% = "negative" ("negative" > correct)
Text: @AmericanAir this is ridiculous! You all really ha..., Prediction:  99.80% = "negative" ("negative" > correct)
Text: @JetBlue's CEO #pilots among ardent fans, Wall Str..., Prediction:  84.67% = "neutral" ("neutral" > correct)
Text: @united I am following - you need to follow me for..., Prediction:  90.00% = "neutral" ("neutral" > correct)
Text: @united funny I paid to check my bag and now fligh..., Prediction:  99.85% = "negative" ("negative" > correct)
Text: @VirginAmerica still waiting to see  @Starryeyes_D..., Prediction:  72.07% = "neutral" ("neutral" > correct)
Text: @SouthwestAir thank you. See u next wednesday in F..., Prediction:  70.88% = "positive" ("positive" > correct)
Text: @USAirways hour 4 at the gate.                    ..., Prediction:  88.03% = "negative" ("negative" > correct)
Text: @united you can't claim "weather" with your hardwo..., Prediction:  98.18% = "negative" ("negative" > correct)
Text: @USAirways you see now that is a lie.  They auto r..., Prediction:  97.97% = "negative" ("negative" > correct)
Text: @usairways  my flight has been Cancelled Flightled..., Prediction:  99.78% = "negative" ("negative" > correct)
Text: @JetBlue Although it wasn't totally the answer I w..., Prediction:  55.87% = "positive" ("positive" > correct)
Text: @JetBlue Well, I try! See you soon!! @JayVig      ..., Prediction:  86.12% = "positive" ("neutral" > wrong)
Text: @united  treats service members like crap never fl..., Prediction:  98.79% = "negative" ("negative" > correct)
Text: @JetBlue I spent an hour on the phone with custome..., Prediction:  99.19% = "negative" ("negative" > correct)
Text: @AmericanAir “Inconvenient” is such a convenient w..., Prediction:  71.39% = "negative" ("negative" > correct)
Text: @JetBlue but the 4 hour policy- when I called and ..., Prediction:  88.69% = "negative" ("negative" > correct)
Text: @AmericanAir that's 16+ extra hours of travel time..., Prediction:  99.55% = "negative" ("negative" > correct)
Text: @SouthwestAir Why can we no longer change trips wi..., Prediction:  98.17% = "negative" ("negative" > correct)
Text: @united Understood and thanks! I should have tried..., Prediction:  71.04% = "positive" ("positive" > correct)
Text: @JetBlue me again :) you don't have ABC on Infligh..., Prediction:  62.26% = "neutral" ("neutral" > correct)
Text: @JetBlue she was a phone agent, pls do! Peggy was ..., Prediction:  74.06% = "positive" ("positive" > correct)
Text: @united WTH be honest with your customers.  This b..., Prediction:  97.63% = "negative" ("negative" > correct)
Text: @united thx for checking in. Never got through via..., Prediction:  99.86% = "negative" ("negative" > correct)
Text: @united why would they make me share a room?      ..., Prediction:  65.19% = "negative" ("negative" > correct)
Text: @United - In case you're reading this, UA230, righ..., Prediction:  78.46% = "negative" ("negative" > correct)
Text: @united  #UnitedAirlines  how long will1531 be del..., Prediction:  83.99% = "negative" ("negative" > correct)

Validation of 50 examples: success: 48/50, accuracy: 96.00, Error rate: 0.04
PS C:\Users\Denis\Documents\DAM-Algorismes>
```

Classify amb input

```
What's your opinion about the airline? I hate it!
Your opinion about the airline is negative with a confidence of 48.19%
PS C:\Users\Denis\Documents\DAM-Algorismes>
```

```
  model.load_state_dict(torch.load(config_file['paths']['trained_network'], map_
What's your opinion about the airline? I love it!
Your opinion about the airline is positive with a confidence of 63.86%
PS C:\Users\Denis\Documents\DAM-Algorismes>
```

```
  model.load_state_dict(torch.load(config_file['paths']['trained_network'], map_locati
What's your opinion about the airline? Could be better
Your opinion about the airline is neutral with a confidence of 47.01%
PS C:\Users\Denis\Documents\DAM-Algorismes>
```