# Multiclass Text Classification with

# Logistic Regression Implemented with PyTorch and CE Loss

First, we will do some initialization.

In [ ]:
```
import random
import torch
import numpy as np
import pandas as pd
from tqdm.notebook import tqdm

# enable tqdm in pandas
tqdm.pandas()

# set to True to use the gpu (if there is one available)
use_gpu = True

# select device
# Se selecciona el gpu envés del cpu para el procesamiento del código
device = torch.device('cuda' if use_gpu and torch.cuda.is_available() else 'cpu')
print(f'device: {device.type}')

# random seed
seed = 1234

# set random seed
if seed is not None:
    print(f'random seed: {seed}')
    random.seed(seed)
    np.random.seed(seed)
    torch.manual_seed(seed)
```

```
device: cpu
random seed: 1234
```

We will be using the AG's News Topic Classification Dataset. It is stored in two CSV files: `train.csv` and `test.csv`, as well as a `classes.txt` that stores the labels of the classes to predict.

First, we will load the training dataset using pandas and take a quick look at how the data.

In [ ]:
```
train_df = pd.read_csv('/kaggle/input/ag-news/train.csv', header=None)
train_df.columns = ['class index', 'title', 'description']
train_df = train_df.sample(frac=0.8,random_state=42)
train_df
```

Out[ ]:

| | class index | title | description |
|---|---|---|---|
| **71787** | 3 | BBC set for major shake-up, claims newspaper | London - The British Broadcasting Corporation,... |
| **67218** | 3 | Marsh averts cash crunch | Embattled insurance broker #39;s banks agree t... |
| **54066** | 2 | Jeter, Yankees Look to Take Control (AP) | AP - Derek Jeter turned a season that started ... |
| **7168** | 4 | Flying the Sun to Safety | When the Genesis capsule comes back to Earth w... |
| **29618** | 3 | Stocks Seen Flat as Nortel and Oil Weigh | NEW YORK (Reuters) - U.S. stocks were set to ... |
| **...** | ... | ... | ... |
| **59228** | 4 | Investors Flock to Web Networking Sites | Internet whiz kids Marc Andreessen, Josh Kopel... |
| **61417** | 3 | Samsung Electric Quarterly Profit Up | Samsung Electronics Co. Ltd. #39;s (005930.KS:... |
| **20703** | 3 | Coeur Still Committed to Wheaton Deal | Coeur d #39;Alene Mines Corp. said Tuesday tha... |
| **40626** | 3 | Clouds on horizon for low-cost airlines | NEW YORK -- As larger US airlines suffer growi... |
| **25059** | 2 | Furcal issues apology for DUI arrest, returns ... | NAMES Atlanta Braves shortstop Rafael Furcal r... |

96000 rows × 3 columns

The dataset consists of 120,000 examples, each consisting of a class index, a title, and a description. The class labels are distributed in a separated file. We will add the labels to the dataset so that we can interpret the data more easily. Note that the label indexes are one-based, so we need to subtract one to retrieve them from the list.

In [ ]:
```
# Se asigna la clase correspondiente a cada titulo y descripcióncon base al indice de clase
labels = open('/kaggle/input/ag-news/classes.txt').read().splitlines()
classes = train_df['class index'].map(lambda i: labels[i-1])
train_df.insert(1, 'class', classes)
train_df
```

Out[ ]:

| | class index | class | title | description |
|---|---|---|---|---|
| **71787** | 3 | Business | BBC set for major shake-up, claims newspaper | London - The British Broadcasting Corporation,... |
| **67218** | 3 | Business | Marsh averts cash crunch | Embattled insurance broker #39;s banks agree t... |
| **54066** | 2 | Sports | Jeter, Yankees Look to Take Control (AP) | AP - Derek Jeter turned a season that started ... |
| **7168** | 4 | Sci/Tech | Flying the Sun to Safety | When the Genesis capsule comes back to Earth w... |
| **29618** | 3 | Business | Stocks Seen Flat as Nortel and Oil Weigh | NEW YORK (Reuters) - U.S. stocks were set to ... |
| **...** | ... | ... | ... | ... |
| **59228** | 4 | Sci/Tech | Investors Flock to Web Networking Sites | Internet whiz kids Marc Andreessen, Josh Kopel... |
| **61417** | 3 | Business | Samsung Electric Quarterly Profit Up | Samsung Electronics Co. Ltd. #39;s (005930.KS:... |
| **20703** | 3 | Business | Coeur Still Committed to Wheaton Deal | Coeur d #39;Alene Mines Corp. said Tuesday tha... |
| **40626** | 3 | Business | Clouds on horizon for low-cost airlines | NEW YORK -- As larger US airlines suffer growi... |
| **25059** | 2 | Sports | Furcal issues apology for DUI arrest, returns ... | NAMES Atlanta Braves shortstop Rafael Furcal r... |

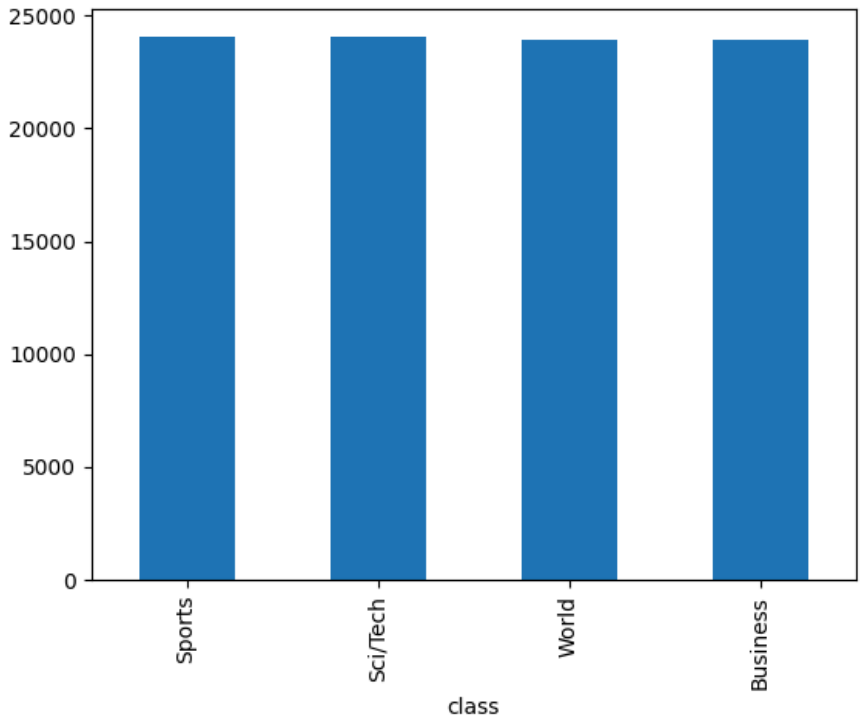
96000 rows × 4 columns

Let's inspect how balanced our examples are by using a bar plot.

In [ ]:
```
pd.value_counts(train_df['class']).plot.bar() # Verificar que todas las clases tengan la misma cantidad de datos para evitar
# sesgos y mejorar el resultado del modelo
```

```
/tmp/ipykernel_30/68118226.py:1: FutureWarning: pandas.value_counts is deprecated and will be removed in a future version. Use pd.
Series(obj).value_counts() instead.
  pd.value_counts(train_df['class']).plot.bar() # Verificar que todas las clases tengan la misma cantidad de datos
```

Out[ ]:
```
<Axes: xlabel='class'>
```




The classes are evenly distributed. That's great!

However, the text contains some spurious backslashes in some parts of the text. They are meant to represent newlines in the original text. An example can be seen below, between the words "dwindling" and "band".

In [ ]:
```
print(train_df.loc[0, 'description'])
```

```
Reuters - Short-sellers, Wall Street's dwindling\band of ultra-cynics, are seeing green again.
```

We will replace the backslashes with spaces on the whole column using pandas replace method.

In [ ]:
```
# Creamos una nueva columna que es la union entre el titulo y la descripción,
# remplazamos \ que representan salto de lineas a espacios para mejorar la
# tokenización de las oraciones así como pasar a minusculas todas las palabras
title = train_df['title'].str.lower()
descr = train_df['description'].str.lower()
text = title + " " + descr
train_df['text'] = text.str.replace('\\', ' ', regex=False)
train_df
```

Out[ ]:

| | class index | class | title | description | text |
|---|---|---|---|---|---|
| **71787** | 3 | Business | BBC set for major shake-up, claims newspaper | London - The British Broadcasting Corporation,... | bbc set for major shake-up, claims newspaper l... |
| **67218** | 3 | Business | Marsh averts cash crunch | Embattled insurance broker #39;s banks agree t... | marsh averts cash crunch embattled insurance b... |
| **54066** | 2 | Sports | Jeter, Yankees Look to Take Control (AP) | AP - Derek Jeter turned a season that started ... | jeter, yankees look to take control (ap) ap - ... |
| **7168** | 4 | Sci/Tech | Flying the Sun to Safety | When the Genesis capsule comes back to Earth w... | flying the sun to safety when the genesis caps... |
| **29618** | 3 | Business | Stocks Seen Flat as Nortel and Oil Weigh | NEW YORK (Reuters) - U.S. stocks were set to ... | stocks seen flat as nortel and oil weigh new ... |
| **...** | ... | ... | ... | ... | ... |
| **59228** | 4 | Sci/Tech | Investors Flock to Web Networking Sites | Internet whiz kids Marc Andreessen, Josh Kopel... | investors flock to web networking sites intern... |
| **61417** | 3 | Business | Samsung Electric Quarterly Profit Up | Samsung Electronics Co. Ltd. #39;s (005930.KS:... | samsung electric quarterly profit up samsung e... |
| **20703** | 3 | Business | Coeur Still Committed to Wheaton Deal | Coeur d #39;Alene Mines Corp. said Tuesday tha... | coeur still committed to wheaton deal coeur d ... |
| **40626** | 3 | Business | Clouds on horizon for low-cost airlines | NEW YORK -- As larger US airlines suffer growi... | clouds on horizon for low-cost airlines new yo... |
| **25059** | 2 | Sports | Furcal issues apology for DUI arrest, returns ... | NAMES Atlanta Braves shortstop Rafael Furcal r... | furcal issues apology for dui arrest, returns ... |

96000 rows × 5 columns

Now we will proceed to tokenize the title and description columns using NLTK's word_tokenize(). We will add a new column to our dataframe with the list of tokens.

In [ ]:
```
from nltk.tokenize import word_tokenize

# Se tokenizan tanto el titulo como la descripción y se crean los vectores de los tokens
train_df['tokens'] = train_df['text'].progress_map(word_tokenize)
train_df
```

```
  0%|          | 0/96000 [00:00<?, ?it/s]
```

Out[ ]:

| | class index | class | title | description | text | tokens |
|---|---|---|---|---|---|---|
| **71787** | 3 | Business | BBC set for major shake-up, claims newspaper | London - The British Broadcasting Corporation,... | bbc set for major shake-up, claims newspaper l... | [bbc, set, for, major, shake-up, ,, claims, ne... |
| **67218** | 3 | Business | Marsh averts cash crunch | Embattled insurance broker #39;s banks agree t... | marsh averts cash crunch embattled insurance b... | [marsh, averts, cash, crunch, embattled, insur... |
| **54066** | 2 | Sports | Jeter, Yankees Look to Take Control (AP) | AP - Derek Jeter turned a season that started ... | jeter, yankees look to take control (ap) ap - ... | [jeter, ,, yankees, look, to, take, control, (... |
| **7168** | 4 | Sci/Tech | Flying the Sun to Safety | When the Genesis capsule comes back to Earth w... | flying the sun to safety when the genesis caps... | [flying, the, sun, to, safety, when, the, gene... |
| **29618** | 3 | Business | Stocks Seen Flat as Nortel and Oil Weigh | NEW YORK (Reuters) - U.S. stocks were set to ... | stocks seen flat as nortel and oil weigh new ... | [stocks, seen, flat, as, nortel, and, oil, wei... |
| **...** | ... | ... | ... | ... | ... | ... |
| **59228** | 4 | Sci/Tech | Investors Flock to Web Networking Sites | Internet whiz kids Marc Andreessen, Josh Kopel... | investors flock to web networking sites intern... | [investors, flock, to, web, networking, sites,... |
| **61417** | 3 | Business | Samsung Electric Quarterly Profit Up | Samsung Electronics Co. Ltd. #39;s (005930.KS:... | samsung electric quarterly profit up samsung e... | [samsung, electric, quarterly, profit, up, sam... |
| **20703** | 3 | Business | Coeur Still Committed to Wheaton Deal | Coeur d #39;Alene Mines Corp. said Tuesday tha... | coeur still committed to wheaton deal coeur d ... | [coeur, still, committed, to, wheaton, deal, c... |
| **40626** | 3 | Business | Clouds on horizon for low-cost airlines | NEW YORK -- As larger US airlines suffer growi... | clouds on horizon for low-cost airlines new yo... | [clouds, on, horizon, for, low-cost, airlines,... |
| **25059** | 2 | Sports | Furcal issues apology for DUI arrest, returns ... | NAMES Atlanta Braves shortstop Rafael Furcal r... | furcal issues apology for dui arrest, returns ... | [furcal, issues, apology, for, dui, arrest, ,,... |

96000 rows × 6 columns

Now we will create a vocabulary from the training data. We will only keep the terms that repeat beyond some threshold established below.

In [ ]:
```
threshold = 10
tokens = train_df['tokens'].explode().value_counts() # Se cuenta la cantidad de repetición de las palabras
tokens = tokens[tokens > threshold] # Se crea un vocabulario solo con palabras que se repitan más de 10 veces
id_to_token = ['[UNK]'] + tokens.index.tolist() # Los ids se convierten a tokens
token_to_id = {w:i for i,w in enumerate(id_to_token)} # Se enumeran los ids convertidos previamente
vocabulary_size = len(id_to_token)
print(f'vocabulary size: {vocabulary_size:,}')
```

```
vocabulary size: 17,430
```

In [ ]:
```
from collections import defaultdict

# Se crea una nueva columna que contiene el id del token y se cuenta las veces que aparece dicho id en la tupla
def make_feature_vector(tokens, unk_id=0):
    vector = defaultdict(int)
    for t in tokens:
        i = token_to_id.get(t, unk_id)
        vector[i] += 1
    return vector

train_df['features'] = train_df['tokens'].progress_map(make_feature_vector) # Se crea un vector de
# caracteristicas basadas en las repeticiones y ids
train_df
```

```
  0%|          | 0/96000 [00:00<?, ?it/s]
```

Out[ ]:

| | class index | class | title | description | text | tokens | features |
|---|---|---|---|---|---|---|---|
| **71787** | 3 | Business | BBC set for major shake-up, claims newspaper | London - The British Broadcasting Corporation,... | bbc set for major shake-up, claims newspaper l... | [bbc, set, for, major, shake-up, ,, claims, ne... | {2455: 1, 167: 1, 11: 1, 200: 1, 6792: 2, 2: 5... |
| **67218** | 3 | Business | Marsh averts cash crunch | Embattled insurance broker #39;s banks agree t... | marsh averts cash crunch embattled insurance b... | [marsh, averts, cash, crunch, embattled, insur... | {1944: 2, 0: 2, 724: 1, 5110: 1, 2891: 1, 753:... |
| **54066** | 2 | Sports | Jeter, Yankees Look to Take Control (AP) | AP - Derek Jeter turned a season that started ... | jeter, yankees look to take control (ap) ap - ... | [jeter, ,, yankees, look, to, take, control, (... | {6647: 2, 2: 1, 508: 1, 599: 1, 4: 1, 193: 1, ... |
| **7168** | 4 | Sci/Tech | Flying the Sun to Safety | When the Genesis capsule comes back to Earth w... | flying the sun to safety when the genesis caps... | [flying, the, sun, to, safety, when, the, gene... | {2603: 1, 1: 4, 415: 2, 4: 3, 1061: 1, 96: 1, ... |
| **29618** | 3 | Business | Stocks Seen Flat as Nortel and Oil Weigh | NEW YORK (Reuters) - U.S. stocks were set to ... | stocks seen flat as nortel and oil weigh new ... | [stocks, seen, flat, as, nortel, and, oil, wei... | {158: 2, 646: 1, 1523: 1, 21: 1, 2036: 2, 9: 1... |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **59228** | 4 | Sci/Tech | Investors Flock to Web Networking Sites | Internet whiz kids Marc Andreessen, Josh Kopel... | investors flock to web networking sites intern... | [investors, flock, to, web, networking, sites,... | {366: 1, 8481: 1, 4: 1, 227: 1, 2620: 1, 992: ... |
| **61417** | 3 | Business | Samsung Electric Quarterly Profit Up | Samsung Electronics Co. Ltd. #39;s (005930.KS:... | samsung electric quarterly profit up samsung e... | [samsung, electric, quarterly, profit, up, sam... | {1744: 2, 2606: 1, 536: 2, 154: 2, 51: 1, 927:... |
| **20703** | 3 | Business | Coeur Still Committed to Wheaton Deal | Coeur d #39;Alene Mines Corp. said Tuesday tha... | coeur still committed to wheaton deal coeur d ... | [coeur, still, committed, to, wheaton, deal, c... | {0: 3, 239: 1, 3350: 2, 4: 2, 9744: 2, 130: 1,... |
| **40626** | 3 | Business | Clouds on horizon for low-cost airlines | NEW YORK -- As larger US airlines suffer growi... | clouds on horizon for low-cost airlines new yo... | [clouds, on, horizon, for, low-cost, airlines,... | {5550: 1, 10: 1, 7485: 1, 11: 1, 2952: 2, 685:... |
| **25059** | 2 | Sports | Furcal issues apology for DUI arrest, returns ... | NAMES Atlanta Braves shortstop Rafael Furcal r... | furcal issues apology for dui arrest, returns ... | [furcal, issues, apology, for, dui, arrest, ,,... | {9255: 3, 951: 1, 6072: 2, 11: 2, 11991: 2, 15... |

96000 rows × 7 columns

In [ ]:
```
# Se reestructuran los features para crear tensors que funcionen con torch
def make_dense(feats):
    x = np.zeros(vocabulary_size)
    for k,v in feats.items():
        x[k] = v
    return x
# Las columnas se convierten al tipo de dato tensor para ser utilizados correctamente con torch
X_train = np.stack(train_df['features'].progress_map(make_dense))
y_train = train_df['class index'].to_numpy() - 1

X_train = torch.tensor(X_train, dtype=torch.float32)
y_train = torch.tensor(y_train)
```

```
  0%|          | 0/96000 [00:00<?, ?it/s]
```

In [ ]:
```
from torch import nn
from torch import optim

# hyperparameters
lr = 1.0
n_epochs = 5
n_examples = X_train.shape[0]
n_feats = X_train.shape[1]
n_classes = len(labels)

# initialize the model, loss function, optimizer, and data-loader
model = nn.Linear(n_feats, n_classes).to(device) # Se define un modelo lineal
loss_func = nn.CrossEntropyLoss() # Se elige la función de perdida "CrossEntropy" por
# ser un problema de clasificación multiclase
optimizer = optim.SGD(model.parameters(), lr=lr) # Se utiliza el optimizador Stochastic Gradient
# Descent para ajustar clasificadores lineales

# train the model
indices = np.arange(n_examples)

# Se entrena el modelo en un bucle con iteraciones definidas con el epoch
for epoch in range(n_epochs):
    np.random.shuffle(indices)
    for i in tqdm(indices, desc=f'epoch {epoch+1}'):
        # clear gradients
        model.zero_grad()
        # send datum to right device
        x = X_train[i].unsqueeze(0).to(device) # Se crea  una nueva dimensión para que sea
        # aceptado por el modelo y se manda al device definido que es el GPU
        y_true = y_train[i].unsqueeze(0).to(device)
        # predict label scores
        y_pred = model(x) # Se predicen labels para validación y comparación para obtener
        # el valor de la funcion de perdida
        # compute loss
```

```
        loss = loss_func(y_pred, y_true)
        # backpropagate
        loss.backward()
        # optimize model parameters
        optimizer.step() # Con base en los resultados de la iteración se mejoran los parametros iniciales del modelo
```

```
epoch 1:   0%|          | 0/96000 [00:00<?, ?it/s]
epoch 2:   0%|          | 0/96000 [00:00<?, ?it/s]
epoch 3:   0%|          | 0/96000 [00:00<?, ?it/s]
epoch 4:   0%|          | 0/96000 [00:00<?, ?it/s]
epoch 5:   0%|          | 0/96000 [00:00<?, ?it/s]
```

Next, we evaluate on the test dataset

In [ ]:
```
# repeat all preprocessing done above, this time on the test set
# Se repite el mismo procedimiento de limpieza, tokenización y creación de tensor
test_df = pd.read_csv('/kaggle/input/ag-news/test.csv', header=None)
test_df = test_df.sample(frac=0.7,random_state=42)
test_df.columns = ['class index', 'title', 'description']
test_df['text'] = test_df['title'].str.lower() + " " + test_df['description'].str.lower()
test_df['text'] = test_df['text'].str.replace('\\', ' ', regex=False)
test_df['tokens'] = test_df['text'].progress_map(word_tokenize)
test_df['features'] = test_df['tokens'].progress_map(make_feature_vector)

X_test = np.stack(test_df['features'].progress_map(make_dense))
y_test = test_df['class index'].to_numpy() - 1
X_test = torch.tensor(X_test, dtype=torch.float32)
y_test = torch.tensor(y_test)
```

```
  0%|          | 0/5320 [00:00<?, ?it/s]
  0%|          | 0/5320 [00:00<?, ?it/s]
  0%|          | 0/5320 [00:00<?, ?it/s]
```

In [ ]:
```
from sklearn.metrics import classification_report

# set model to evaluation mode
model.eval()

# don't store gradients
with torch.no_grad():
    X_test = X_test.to(device)
    y_pred = torch.argmax(model(X_test), dim=1)
    y_pred = y_pred.cpu().numpy()
    print(classification_report(y_test, y_pred, target_names=labels))

# Se predice en las distintas categorías y se obtienen las métricas de precisión con los valores reales de pruebas
```

```
              precision    recall  f1-score   support

       World       0.88      0.91      0.89      1330
      Sports       0.90      0.98      0.94      1334
    Business       0.84      0.85      0.84      1314
    Sci/Tech       0.91      0.80      0.85      1342

    accuracy                           0.88      5320
   macro avg       0.88      0.88      0.88      5320
weighted avg       0.88      0.88      0.88      5320
```

In [ ]:
```
from google.colab import drive
drive.mount('/content/drive')
```

In [ ]:
```
# Exportar a HTML
!jupyter nbconvert --to html "/content/drive/MyDrive/Colab Notebooks/notebookd1d7b02f75.ipynb"
```