Multiclass Text Classification with

Logistic Regression Implemented with PyTorch and CE Loss

First, we will do some initialization.

In [3]:
```
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
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 [4]:
```
train_df = pd.read_csv('/kaggle/input/agnews-pytorch-simple-embed-classif-90/AG_NEW
train_df.columns = ['class index', 'title', 'description']

train_df = train_df.sample(frac=0.8, random_state=42)
train_df
```

Out[4]:

	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 [5]:
```
labels = open('/kaggle/input/classes/classes.txt').read().splitlines()
classes = train_df['class index'].map(lambda i: labels[i-1])
train_df.insert(1, 'class', classes)
train_df
```

Out[5]:

	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...

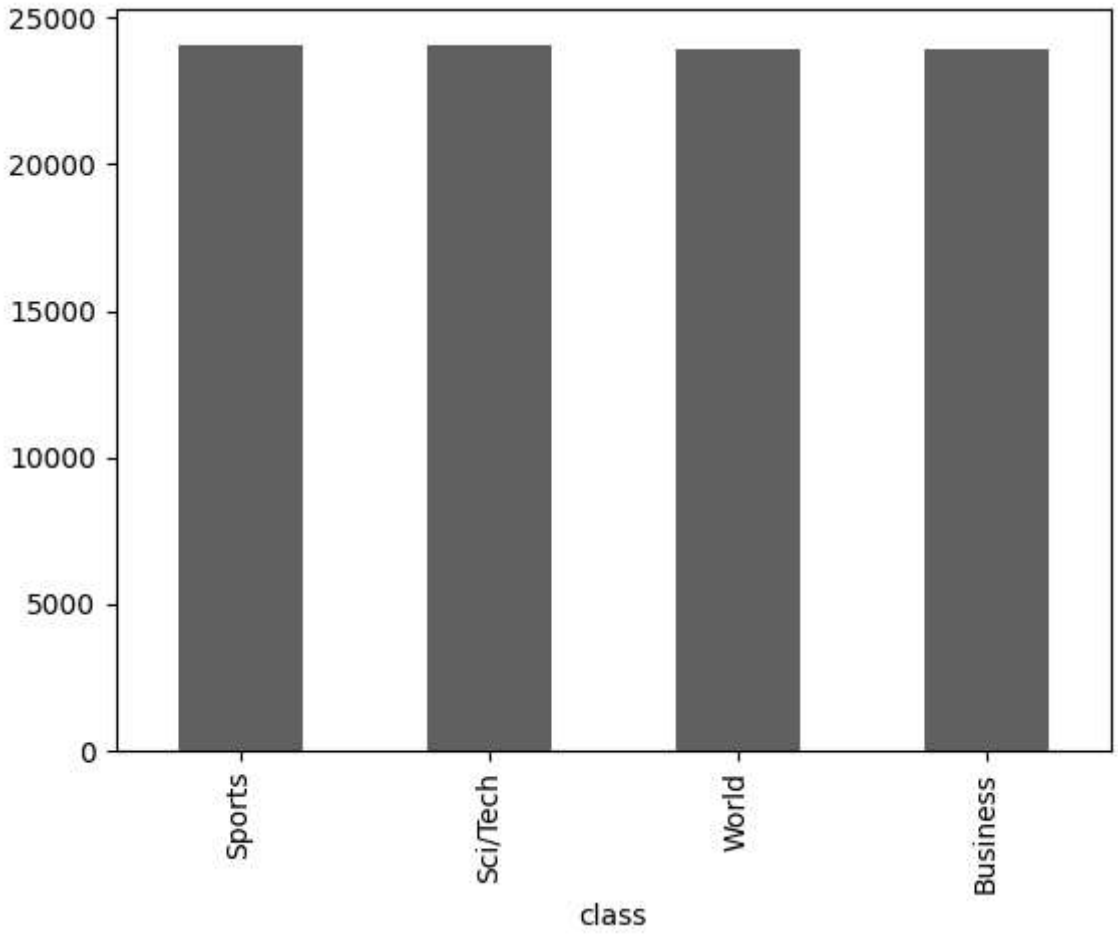
96000 rows × 4 columns

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

In [6]:
```
pd.value_counts(train_df['class']).plot.bar()
```

```
/tmp/ipykernel_124/1245903889.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()
```

Out[6]:
```
<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 [7]:
```
print(train_df.loc[0, 'description'])
```

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

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

In [8]:
```
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[8]:

	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 [9]:
```
from nltk.tokenize import word_tokenize

# Tokenizamos cada texto en la columna 'text' de nuestro DataFrame 'train_df' y gua
# Se utiliza 'progress_map' para aplicar el tokenizador y visualizar una barra de p
train_df['tokens'] = train_df['text'].progress_map(word_tokenize)
```

```
train_df
```

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

Out[9]:

	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 [10]:
```
# Definimos un umbral mínimo de frecuencia para los tokens.
threshold = 10

# Contamos la frecuencia de cada token en la columna 'tokens' del DataFrame 'train_
# Usamos 'explode' para dividir cada lista de tokens en elementos individuales y lu
tokens = train_df['tokens'].explode().value_counts()

# Filtramos los tokens para quedarnos solo con aquellos cuya frecuencia supera el u
tokens = tokens[tokens > threshold]

# Creamos una lista 'id_to_token' que contiene un token especial '[UNK]' y todos lo
id_to_token = ['[UNK]'] + tokens.index.tolist()

# Creamos un diccionario 'token_to_id' que asigna un ID único a cada token, donde l
token_to_id = {w:i for i,w in enumerate(id_to_token)}


vocabulary_size = len(id_to_token)


print(f'vocabulary size: {vocabulary_size:,}')
```

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

In [11]:
```
from collections import defaultdict

# Definimos la función 'make_feature_vector', que toma una lista de tokens y un ide
def make_feature_vector(tokens, unk_id=0):
    # Inicializamos un vector de características como un diccionario que asigna un
    vector = defaultdict(int)

    # Iteramos sobre cada token en la lista de tokens proporcionada.
    for t in tokens:
        # Obtenemos el ID del token correspondiente del diccionario 'token_to_id',
        i = token_to_id.get(t, unk_id)
        # Incrementamos el contador del token en el vector de características.
        vector[i] += 1

    return vector  # Retornamos el vector de características resultante.

# Aplicamos la función 'make_feature_vector' a la columna 'tokens' del DataFrame 't
# Usamos 'progress_map' para mostrar el progreso del procesamiento.
train_df['features'] = train_df['tokens'].progress_map(make_feature_vector)

# Mostramos el DataFrame 'train_df' actualizado con la nueva columna 'features'.
train_df
```

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

Out[11]:

	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,...

	class index	class	title	description	text	tokens	features
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 [12]:
```
def make_dense(feats):
    x = np.zeros(vocabulary_size)

    # Iteramos sobre cada par clave-valor en el diccionario de características.
    for k, v in feats.items():
        # Asignamos el valor del contador al índice correspondiente en el vector de
        x[k] = v

    return x  # Retornamos el vector denso resultante.

# Aplicamos la función 'make_dense' a la columna 'features' del DataFrame 'train_df
# y apilamos los resultados en un array 2D usando 'np.stack'.
X_train = np.stack(train_df['features'].progress_map(make_dense))

# Convertimos la columna 'class index' del DataFrame a un array de NumPy y restamos
y_train = train_df['class index'].to_numpy() - 1

# Convertimos 'X_train' y 'y_train' a tensores de PyTorch.
# Especificamos que 'X_train' será de tipo float32.
X_train = torch.tensor(X_train, dtype=torch.float32)
y_train = torch.tensor(y_train)
```

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

In [13]:
```
from torch import nn
from torch import optim

# Definimos los hiperparámetros para el entrenamiento del modelo.
lr = 1.0  # Tasa de aprendizaje.
n_epochs = 5  # Número de épocas para entrenar el modelo.
n_examples = X_train.shape[0]  # Número total de ejemplos en el conjunto de entrena
n_feats = X_train.shape[1]  # Número de características (dimensiones) en cada ejemp
n_classes = len(labels)  # Número de clases en nuestro conjunto de datos.

# Inicializamos el modelo, la función de pérdida, el optimizador y el cargador de d
# En este caso, usamos una capa lineal simple como modelo.
model = nn.Linear(n_feats, n_classes).to(device)
```

```
# Definimos la función de pérdida como entropía cruzada, adecuada para problemas de
loss_func = nn.CrossEntropyLoss()
# Usamos el optimizador SGD (Stochastic Gradient Descent) para actualizar los parám
optimizer = optim.SGD(model.parameters(), lr=lr)

# Comenzamos el proceso de entrenamiento del modelo.
indices = np.arange(n_examples)  # Creamos un array de índices para los ejemplos.
for epoch in range(n_epochs):  # Iteramos a través de cada época de entrenamiento.
    np.random.shuffle(indices)  # Mezclamos los índices en cada época para garantiz
    for i in tqdm(indices, desc=f'epoch {epoch+1}'):  # Iteramos sobre los índices
        # Limpiamos los gradientes acumulados de la iteración anterior.
        model.zero_grad()

        # Enviamos el dato actual al dispositivo correcto (CPU o GPU).
        x = X_train[i].unsqueeze(0).to(device)  # Añadimos una dimensión para repre
        y_true = y_train[i].unsqueeze(0).to(device)  # También hacemos lo mismo par

        # Usamos el modelo para predecir las puntuaciones de las etiquetas.
        y_pred = model(x)

        # Calculamos la pérdida entre las predicciones y la verdad conocida.
        loss = loss_func(y_pred, y_true)

        # Realizamos la retropropagación para calcular los gradientes.
        loss.backward()

        # Actualizamos los parámetros del modelo usando el optimizador.
        optimizer.step()
```

```
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 [15]:
```
# repeat all preprocessing done above, this time on the test set
test_df = pd.read_csv('/kaggle/input/agnews-pytorch-simple-embed-classif-90/AG_NEWS
test_df.columns = ['class index', 'title', 'description']
test_df['text'] = test_df['title'].str.lower() + " " + test_df['description'].str.l
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%|          | 0/7600 [00:00<?, ?it/s]
 0%|          | 0/7600 [00:00<?, ?it/s]
 0%|          | 0/7600 [00:00<?, ?it/s]
```

In [16]:
```
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() # Se usa el CPU
    print(classification_report(y_test, y_pred, target_names=labels))
```

```
              precision    recall  f1-score   support

       World       0.88      0.91      0.89      1900
      Sports       0.90      0.98      0.94      1900
    Business       0.84      0.85      0.85      1900
    Sci/Tech       0.90      0.78      0.84      1900

    accuracy                           0.88      7600
   macro avg       0.88      0.88      0.88      7600
weighted avg       0.88      0.88      0.88      7600
```