Multiclass Text Classification with

Logistic Regression Implemented with PyTorch and CE Loss

First, we will do some initialization.

In [1]:
```
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: cuda
random seed: 1234
```

Este código habilita tqdm para visualizar barras de progreso en operaciones de Pandas. Luego, define si se usará una GPU (use_gpu=True) y selecciona el dispositivo adecuado (cuda para GPU o cpu en caso contrario) mediante torch.device. También establece una semilla aleatoria (seed = 1234) para asegurar que siempre se usen los mismos datos para que las ejecuciones sean consistentes.

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 [2]:
```
train_df = pd.read_csv('/kaggle/input/train-csv/train.csv', header=None)
train_df = train_df.sample(frac = 0.8, random_state = 42)
```

```
train_df.columns = ['class index', 'title', 'description']
train_df
```

Out[2]:

	class index	title	description
71788	3	BBC set for major shake-up, claims newspaper	London - The British Broadcasting Corporation,...
67218	4	Taking Microsoft for a spin?	The software juggernaut that conquered the des...
54066	3	September sales at Target stores beat retail a...	MINNEAPOLIS - While other retailers struggled ...
7168	4	Macromedia launches Flex Builder	Macromedia this week will ship Flex Builder, w...
29618	1	Rocket lands near Afghan school as President K...	AFP - A rocket landed near a school in southea...
...	...	...	...
59228	4	Technical Problems Subside at PayPal	Most members of the online payment service Pay...
61417	3	Shoppers Spring Back to Life in September	Shoppers got their buying groove back last mon...
20703	3	UPDATE 1-Yellow Roadway raises 3rd-qtr profit ...	Yellow Roadway Corp. (YELL.O: Quote, Profile, ...
40626	3	Next to digital IDs, passwords look lame	How big is your key ring? There are the house ...
25059	2	Prime-time Eagles	They opened their season Sept. 2 in the smalle...

96001 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 [3]:

```
labels = open('/kaggle/input/classes-txt/classes.txt').read().splitlines()
train_df = train_df.drop(0).reset_index(drop=True)
train_df['class index'] = train_df['class index'].astype(int)
classes = train_df['class index'].map(lambda i: labels[i-1])
train_df.insert(1, 'class', classes)
train_df
```

Out[3]:

	class index	class	title	description
0	3	Business	BBC set for major shake-up, claims newspaper	London - The British Broadcasting Corporation,...
1	4	Sci/Tech	Taking Microsoft for a spin?	The software juggernaut that conquered the des...
2	3	Business	September sales at Target stores beat retail a...	MINNEAPOLIS - While other retailers struggled ...
3	4	Sci/Tech	Macromedia launches Flex Builder	Macromedia this week will ship Flex Builder, w...
4	1	World	Rocket lands near Afghan school as President K...	AFP - A rocket landed near a school in southea...
...	...	...	...	...
95995	4	Sci/Tech	Technical Problems Subside at PayPal	Most members of the online payment service Pay...
95996	3	Business	Shoppers Spring Back to Life in September	Shoppers got their buying groove back last mon...
95997	3	Business	UPDATE 1-Yellow Roadway raises 3rd-qtr profit ...	Yellow Roadway Corp. (YELL.O: Quote, Profile, ...
95998	3	Business	Next to digital IDs, passwords look lame	How big is your key ring? There are the house ...
95999	2	Sports	Prime-time Eagles	They opened their season Sept. 2 in the smalle...

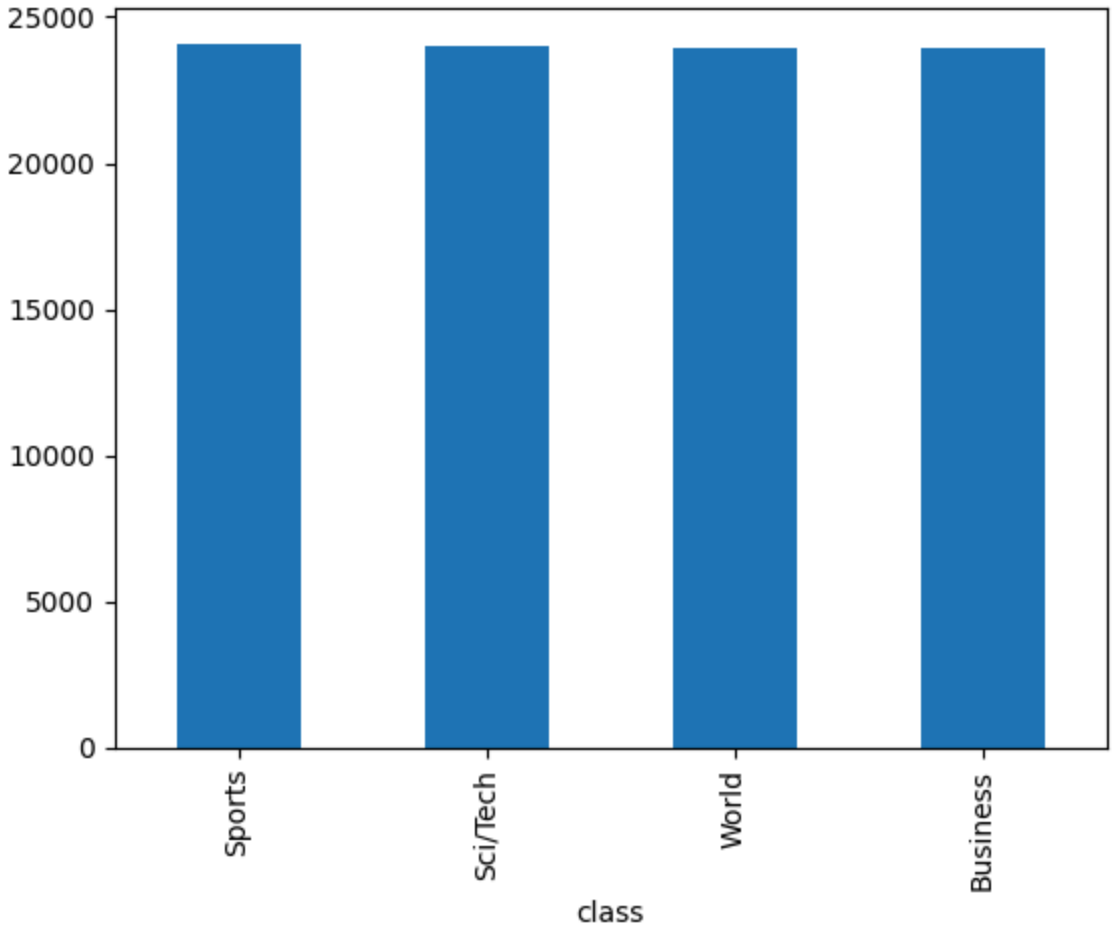
96000 rows × 4 columns

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

In [4]:
```
pd.value_counts(train_df['class']).plot.bar()
```

```
/tmp/ipykernel_30/1245903889.py:1: FutureWarning: pandas.value_counts is deprecated a
nd will be removed in a future version. Use pd.Series(obj).value_counts() instead.
  pd.value_counts(train_df['class']).plot.bar()
```

Out[4]:
```
<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 [5]:
```
print(train_df.loc[0, 'description'])
```

```
London - The British Broadcasting Corporation, the world #39;s biggest public broadca
ster, is to cut almost a quarter of its 28 000-strong workforce, in the biggest shake
-up in its 82-year history, The Times newspaper in London said on Monday.
```

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

In [6]:
```
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[6]:

	class index	class	title	description	text
0	3	Business	BBC set for major shake-up, claims newspaper	London - The British Broadcasting Corporation,...	bbc set for major shake-up, claims newspaper l...
1	4	Sci/Tech	Taking Microsoft for a spin?	The software juggernaut that conquered the des...	taking microsoft for a spin? the software jugg...
2	3	Business	September sales at Target stores beat retail a...	MINNEAPOLIS - While other retailers struggled ...	september sales at target stores beat retail a...
3	4	Sci/Tech	Macromedia launches Flex Builder	Macromedia this week will ship Flex Builder, w...	macromedia launches flex builder macromedia th...
4	1	World	Rocket lands near Afghan school as President K...	AFP - A rocket landed near a school in southea...	rocket lands near afghan school as president k...
...	...	...	...	...	...
95995	4	Sci/Tech	Technical Problems Subside at PayPal	Most members of the online payment service Pay...	technical problems subside at paypal most memb...
95996	3	Business	Shoppers Spring Back to Life in September	Shoppers got their buying groove back last mon...	shoppers spring back to life in september shop...
95997	3	Business	UPDATE 1-Yellow Roadway raises 3rd-qtr profit ...	Yellow Roadway Corp. (YELL.O: Quote, Profile, ...	update 1-yellow roadway raises 3rd-qtr profit ...
95998	3	Business	Next to digital IDs, passwords look lame	How big is your key ring? There are the house ...	next to digital ids, passwords look lame how b...
95999	2	Sports	Prime-time Eagles	They opened their season Sept. 2 in the smalle...	prime-time eagles they opened their season sep...

96000 rows × 5 columns

Crea una nueva columna text que combina el título y la descripción en minúsculas, ayudando a normalizar los datos para la tarea de procesamiento de lenguaje natural.

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 [7]:
```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /usr/share/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[7]:
```
True
```

In [8]:
```
from nltk.tokenize import word_tokenize

train_df['tokens'] = train_df['text'].progress_map(word_tokenize)
train_df
```

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

Out[8]:

	class index	class	title	description	text	tokens
0	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...
1	4	Sci/Tech	Taking Microsoft for a spin?	The software juggernaut that conquered the des...	taking microsoft for a spin? the software jugg...	[taking, microsoft, for, a, spin, ?, the, soft...
2	3	Business	September sales at Target stores beat retail a...	MINNEAPOLIS - While other retailers struggled ...	september sales at target stores beat retail a...	[september, sales, at, target, stores, beat, r...
3	4	Sci/Tech	Macromedia launches Flex Builder	Macromedia this week will ship Flex Builder, w...	macromedia launches flex builder macromedia th...	[macromedia, launches, flex, builder, macromed...
4	1	World	Rocket lands near Afghan school as President K...	AFP - A rocket landed near a school in southea...	rocket lands near afghan school as president k...	[rocket, lands, near, afghan, school, as, pres...
...	...	...	...	...	...	...
95995	4	Sci/Tech	Technical Problems Subside at PayPal	Most members of the online payment service Pay...	technical problems subside at paypal most memb...	[technical, problems, subside, at, paypal, mos...
95996	3	Business	Shoppers Spring Back to Life in September	Shoppers got their buying groove back last mon...	shoppers spring back to life in september shop...	[shoppers, spring, back, to, life, in, septemb...
95997	3	Business	UPDATE 1-Yellow Roadway raises 3rd-qtr profit ...	Yellow Roadway Corp. (YELL.O: Quote, Profile, ...	update 1-yellow roadway raises 3rd-qtr profit ...	[update, 1-yellow, roadway, raises, 3rd-qtr, p...
95998	3	Business	Next to digital IDs, passwords look lame	How big is your key ring? There are the house ...	next to digital ids, passwords look lame how b...	[next, to, digital, ids, ,, passwords, look, l...
95999	2	Sports	Prime-time Eagles	They opened their season Sept. 2 in the smalle...	prime-time eagles they opened their season sep...	[prime-time, eagles, they, opened, their, seas...

96000 rows × 6 columns

Este código usa word_tokenize para dividir el texto en palabras, que serían los tokens. Esto se aplica a cada fila de la columna text en train_df, creando una nueva columna

tokens que contiene listas de palabras tokenizadas para cada texto. Igualmente, la función progress_map muestra el progreso de la operación gracias a tqdm, lo cual es útil para saber cuánto tiempo resta en el procesamiento de datos largos.

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

In [9]:
```
threshold = 10
tokens = train_df['tokens'].explode().value_counts()
tokens = tokens[tokens > threshold]
id_to_token = ['[UNK]'] + tokens.index.tolist()
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: 17,436
```

Este código crea un vocabulario de tokens que aparecen más de 10 veces en train_df. Primero, value_counts() cuenta la frecuencia de cada token y luego se filtran aquellos con frecuencia mayor a threshold. [UNK] se agrega como token desconocido, y id_to_token contiene todos los tokens restantes. token_to_id es un diccionario que asigna un ID único a cada token. Finalmente, vocabulary_size guarda el tamaño total del vocabulario y se imprime.

In [10]:
```
from collections import defaultdict

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)
train_df
```

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

Out[10]:

	class index	class	title	description	text	tokens	features
0	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...	{2729: 1, 168: 1, 11: 1, 204: 1, 7015: 2, 2: 5...
1	4	Sci/Tech	Taking Microsoft for a spin?	The software juggernaut that conquered the des...	taking microsoft for a spin? the software jugg...	[taking, microsoft, for, a, spin, ?, the, soft...	{612: 1, 84: 1, 11: 1, 5: 1, 4586: 1, 88: 1, 1...
2	3	Business	September sales at Target stores beat retail a...	MINNEAPOLIS - While other retailers struggled ...	september sales at target stores beat retail a...	[september, sales, at, target, stores, beat, r...	{446: 1, 131: 2, 22: 1, 782: 2, 599: 1, 377: 1...
3	4	Sci/Tech	Macromedia launches Flex Builder	Macromedia this week will ship Flex Builder, w...	macromedia launches flex builder macromedia th...	[macromedia, launches, flex, builder, macromed...	{5419: 2, 965: 1, 8376: 3, 7550: 2, 59: 1, 93:...
4	1	World	Rocket lands near Afghan school as President K...	AFP - A rocket landed near a school in southea...	rocket lands near afghan school as president k...	[rocket, lands, near, afghan, school, as, pres...	{1129: 2, 3801: 1, 365: 2, 704: 1, 535: 2, 21:...
...	...	...	...	...	...	...	...
95995	4	Sci/Tech	Technical Problems Subside at PayPal	Most members of the online payment service Pay...	technical problems subside at paypal most memb...	[technical, problems, subside, at, paypal, mos...	{2445: 1, 911: 1, 0: 1, 22: 1, 4009: 2, 147: 1...
95996	3	Business	Shoppers Spring Back to Life in September	Shoppers got their buying groove back last mon...	shoppers spring back to life in september shop...	[shoppers, spring, back, to, life, in, septemb...	{2762: 2, 2649: 1, 119: 2, 4: 1, 486: 1, 7: 1,...
95997	3	Business	UPDATE 1-Yellow Roadway raises 3rd-qtr profit ...	Yellow Roadway Corp. (YELL.O: Quote, Profile, ...	update 1-yellow roadway raises 3rd-qtr profit ...	[update, 1-yellow, roadway, raises, 3rd-qtr, p...	{347: 1, 0: 2, 12962: 2, 1453: 1, 11057: 1, 16...
95998	3	Business	Next to digital IDs, passwords look lame	How big is your key ring? There are the house ...	next to digital ids, passwords look lame how b...	[next, to, digital, ids, ,, passwords, look, l...	{118: 1, 4: 3, 449: 1, 0: 4, 2: 5, 6026: 1, 60...
95999	2	Sports	Prime-time Eagles	They opened their season	prime-time eagles they	[prime-time, eagles, they,	{10794: 1, 1360: 1, 74: 1,

	class index	class	title	description	text	tokens	features
				Sept. 2 in the smalle...	opened their season sep...	opened, their, seas...	1214: 1, 47: 1, 126...

96000 rows × 7 columns

La función make_feature_vector convierte una lista de tokens en un vector de características basado en la frecuencia de cada token. Cada token se mapea a su ID usando token_to_id, o a unk_id si el token no está. El resultado es un diccionario (vector) con los IDs de tokens como claves y sus frecuencias como valores. Luego, train_df['features'] = train_df['tokens'].progress_map(make_feature_vector) aplica esta función a cada fila de tokens, creando la columna features en train_df, con una barra de progreso para mostrar el avance.

In [11]:
```
def make_dense(feats):
    x = np.zeros(vocabulary_size)
    for k,v in feats.items():
        x[k] = v
    return x

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]
```

La función make_dense convierte el diccionario de tokens feats en un vector denso de tamaño vocabulary_size. Inicializa un vector de ceros y asigna las frecuencias de cada token según el índice del vocabulario. X_train se construye aplicando make_dense a cada fila en train_df['features'], creando una matriz con vectores de frecuencias densos. y_train convierte la columna class index en un array, ajustando los índices para comenzar en 0. Finalmente, X_train y y_train se vuelven tensores de PyTorch para poder usarlos en los modelos.

In [12]:
```
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)
loss_func = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(), lr=lr)
```

```
# train the model
indices = np.arange(n_examples)
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)
        y_true = y_train[i].unsqueeze(0).to(device)
        # predict label scores
        y_pred = model(x)
        # compute loss
        loss = loss_func(y_pred, y_true)
        # backpropagate
        loss.backward()
        # optimize model parameters
        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]
```

Este chunk configura y entrena un modelo de clasificación usando Pytorch. Primero define hiperparámetros como la tasa de aprendizaje (lr), el número de épocas (n_epochs), y el tamaño de los datos de entrenamiento (n_examples, n_feats, y n_classes). Después, inicializa el modelo (nn.Linear), la función de pérdida (entropía cruzada, nn.CrossEntropyLoss), y el optimizador (SGD con tasa de aprendizaje lr). Durante cada epoch de entrenamiento, los índices de los ejemplos se mezclan para hacer cada iteracion mas variable. En general el código:

- Limpia los gradientes previos.
- Envía el ejemplo x y su etiqueta y_true al dispositivo adecuado (GPU o CPU).
- Calcula la predicción del modelo (y_pred).
- Calcula la pérdida comparando y_pred con y_true.
- Realiza retropropagación para ajustar los gradientes (loss.backward()).
- Optimiza los parámetros del modelo (optimizer.step()).

Next, we evaluate on the test dataset

In [16]:
```
# repeat all preprocessing done above, this time on the test set
test_df = pd.read_csv('/kaggle/input/test-csv/test.csv', header=None)
test_df.columns = ['class index', 'title', 'description']
test_df['text'] = test_df['title'].str.lower() + " " + test_df['description'].str.lowe
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)

test_df = test_df.drop(index=0).reset_index(drop=True)
test_df['class index'] = pd.to_numeric(test_df['class index'], errors='coerce')

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%|          | 0/7601 [00:00<?, ?it/s]
  0%|          | 0/7601 [00:00<?, ?it/s]
  0%|          | 0/7600 [00:00<?, ?it/s]
```

El código anterior repite todo el procedimiento pasado pero ahora en el dataset de test.

In [17]:
```
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))
```

```
              precision    recall  f1-score   support

       World       0.96      0.79      0.87      1900
      Sports       0.92      0.98      0.95      1900
    Business       0.75      0.91      0.82      1900
    Sci/Tech       0.87      0.79      0.83      1900

    accuracy                           0.86      7600
   macro avg       0.87      0.86      0.86      7600
weighted avg       0.87      0.86      0.86      7600
```

En este código se evalúa el modelo y se puede notar que tuvo un buen desempeño de acuerdo a los resultados que nos dio, como el accuracy alto y también el f1-score.