Juan Pablo Bernal Lafarga - A01742342

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 '
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 bloque de código configura el entorno para realizar experimentos reproducibles en machine learning. Activa barras de progreso en pandas, selecciona la GPU o CPU según disponibilidad, y establece una semilla aleatoria para asegurar que los resultados sean consistentes en cada ejecución.

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/d/kk0105/ag-news/ag_news_csv/train.csv
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
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

Este bloque prepara un subconjunto de datos de entrenamiento para análisis o modelado.

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/d/kk0105/ag-news/ag_news_csv/classes.txt').reac
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
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

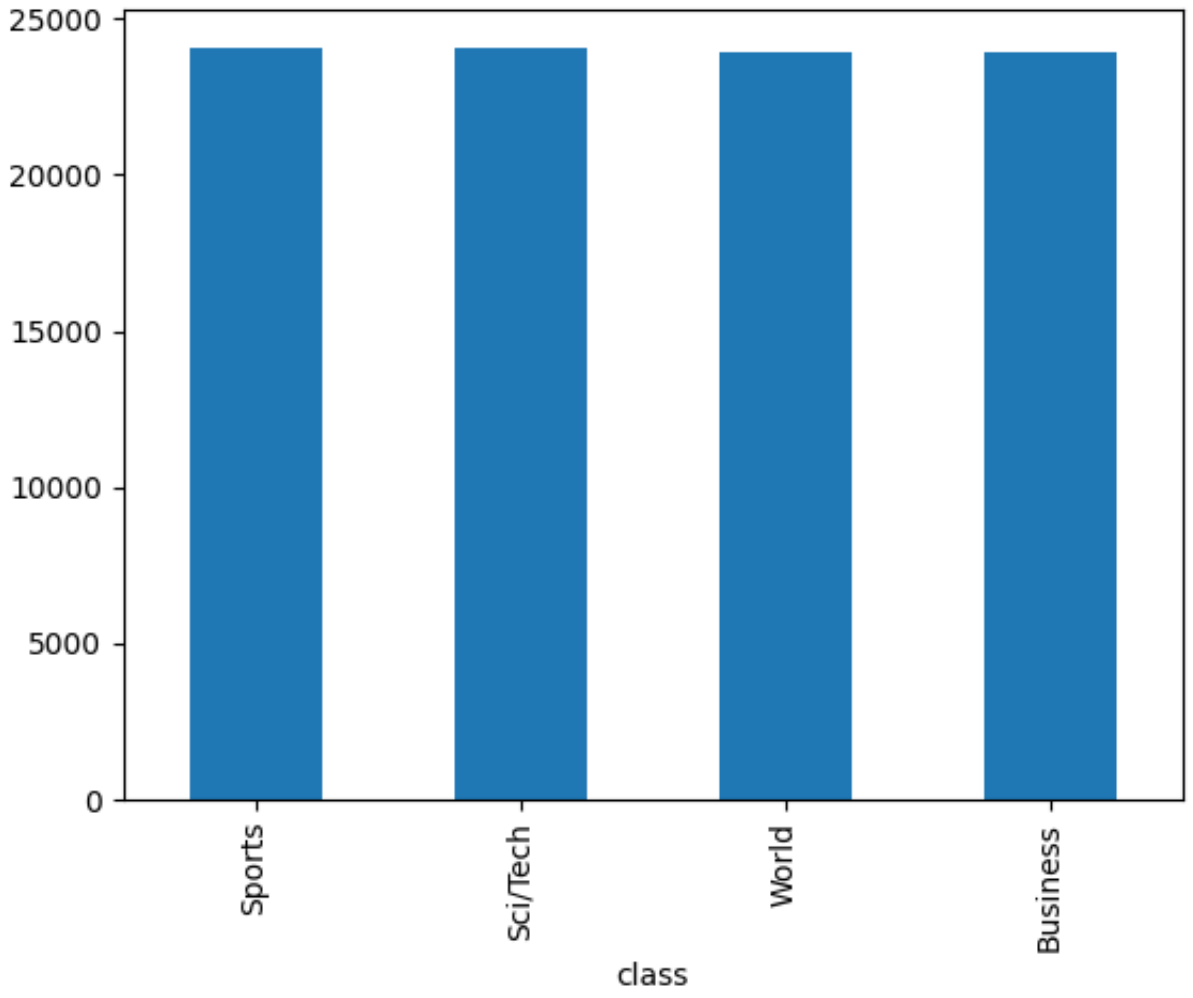
Este bloque de código carga las etiquetas de clase desde un archivo de texto, las asigna a los datos y las inserta en el data frame

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_210/1245903889.py:1: FutureWarning: pandas.value_counts is de
precated and will be removed in a future version. Use pd.Series(obj).value_c
ounts() instead.
  pd.value_counts(train_df['class']).plot.bar()
```

Out[4]: <Axes: xlabel='class'>



Este bloque de código genera una gráfica de barras para una clase del Data Frame

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'])
```

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

Este bloque de código muestra la primera instancia de la clase "description" del DataFrame de entrenamiento.

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
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

Este bloque de código crea una nueva columna en el DataFrame de entrenamiento llamada "text" en el cual se encuentra una fusión de las columnas "title" y "description", con toda la información en minúsculas.

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]:
```
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[7]:

	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...
			Coeur Still	Coeur d	coeur still	[coeur, still,

20703	3	Business	Committed to Wheaton Deal	#39;Alene Mines Corp. said Tuesday tha...	committed to wheaton deal coeur d ...	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

Este bloque de código tokeniza el texto en la columna 'text' del DataFrame usando word_tokenize de la librería nltk y almacena los tokens en una nueva columna.

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

In [8]:
```
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: 17,430
```

Este bloque de código crea un vocabulario basado en la frecuencia de las palabras en el conjunto de datos, con un umbral mínimo, y asigna identificadores únicos a cada token.

In [9]:
```
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[9]:

	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...	{2451: 1, 167: 1, 11: 1, 201: 1, 6778: 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...	{1945: 2, 0: 2, 723: 1, 5100: 1, 2891: 1, 752:...
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, (...	{6670: 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...	{2601: 1, 1: 4, 416: 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, 2035: 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, 8544: 1, 4: 1, 227: 1, 2620: 1, 992: ...
					samsung	[samsung,	{1745: 2,

61417	3	Business	Samsung Electric Quarterly Profit Up	Samsung Electronics Co. Ltd. #39;s (005930.KS:...	electric quarterly profit up samsung e...	electric, quarterly, profit, up, sam...	2597: 1, 536: 2, 154: 2, 51: 1, 926:...
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, 3351: 2, 4: 2, 9726: 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,...	{5532: 1, 10: 1, 7500: 1, 11: 1, 2949: 2, 683:...
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, ,,...	{9372: 3, 951: 1, 6078: 2, 11: 2, 11962: 2, 15...

96000 rows × 7 columns

Este bloque de código crea una representación de "bag of words" para cada conjunto de tokens, contando la frecuencia de cada token en un vector de características (feature vector).

In [10]:
```
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]
```

Este bloque de código convierte los vectores de características dispersos en vectores densos de tamaño fijo, adecuados para usarse en modelos de aprendizaje profundo.

In [11]:
```
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 bloque de código configura y entrena un modelo de regresión logística para clasificación, en el que cada época recorre aleatoriamente los ejemplos de entrenamiento, calcula las predicciones y ajusta los parámetros del modelo mediante retropropagación y optimización estocástica.

Next, we evaluate on the test dataset

In [13]:
```
# repeat all preprocessing done above, this time on the test set
```

```
test_df = pd.read_csv('/kaggle/input/d/kk0105/ag-news/ag_news_csv/test.csv',
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'
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]
```

Este bloque de código realiza todo el proceso de preprocesamiento necesario para el conjunto de prueba, replicando lo que se hizo para el conjunto de entrenamiento. Esto incluye cargar los datos, limpiar y tokenizar el texto, generar vectores de características y finalmente convertir los datos en tensores de PyTorch, preparándolos para la evaluación del modelo.

In [14]:
```
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       0.87      0.91      0.89      1330
      Sports       0.95      0.96      0.95      1334
    Business       0.85      0.86      0.85      1314
    Sci/Tech       0.89      0.84      0.86      1342

    accuracy                           0.89      5320
   macro avg       0.89      0.89      0.89      5320
weighted avg       0.89      0.89      0.89      5320
```

Este bloque de código evalúa el modelo en el conjunto de datos de prueba, calcula las predicciones y genera un informe de clasificación que resume el desempeño del

modelo en términos de precisión, recuperación y F1-score. Esto es fundamental para entender cómo de bien está funcionando el modelo en datos no vistos.

El informe indica que el modelo tiene un buen rendimiento general en el conjunto de prueba, con una precisión, recall y F1-score promedio de alrededor del 89% para todas las clases. No obstante, la precisión y el recall son más altos en la clase "Sports", lo que puede sugerir que el modelo es más efectivo para clasificar instancias en esa categoría. En contraste, la clase "Business" tiene las métricas más bajas, lo que podría indicar que el modelo tiene más dificultades para identificar correctamente las instancias de esa clase.