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: 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.

La razon del porque seleccionamos 70% fue debido a que nos ayuda a prevenir problemas debido a los recursos limitados

In [2]:

```
#Obtenemos la informacion de dataset de train, para poder obtener las clases, a
#un 70% de los datos son usados de entrenamiento
train_df = pd.read_csv('/kaggle/input/agnews-pytorch-simple-embed-classif-90/AG
train_df.columns = ['class index', 'title', 'description']
train_df = train_df.sample(frac=0.7,random_state=42)
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 ...
...	...	...	...
53857	1	FDA Accused of Silencing Vioxx Warnings	WASHINGTON - The Food and Drug Administration ...
111476	2	Buckeyes won #39;t play in NCAA or NIT tourneys	COLUMBUS, Ohio Ohio State has sanctioned its m...
6343	3	Rate hikes by Fed work in two ways	If you #39;ve noticed that the price of everyt...
20736	4	NASA Administrator Offers Support for Kennedy ...	The following is a statement from NASA Adminis...
34378	2	Twins make it 3 straight	The Minnesota Twins clinched on a bus in 1991....

84000 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.

El asignar las etiuqetas ayudan a poder interpretar los resultados fiinales de una manera mas sencilla

In [3]:

```
#Obtiene los titulos de las classes, los cuales se encuentran en el documento
labels = open('/kaggle/input/newnasmes/classes.txt').read().splitlines()
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 ...
...	...	...	...	...
53857	1	World	FDA Accused of Silencing Vioxx Warnings	WASHINGTON - The Food and Drug Administration ...
111476	2	Sports	Buckeyes won #39;t play in NCAA or NIT tourneys	COLUMBUS, Ohio Ohio State has sanctioned its m...
6343	3	Business	Rate hikes by Fed work in two ways	If you #39;ve noticed that the price of everyt...
20736	4	Sci/Tech	NASA Administrator Offers Support for Kennedy ...	The following is a statement from NASA Adminis...
34378	2	Sports	Twins make it 3 straight	The Minnesota Twins clinched on a bus in 1991....

84000 rows × 4 columns

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

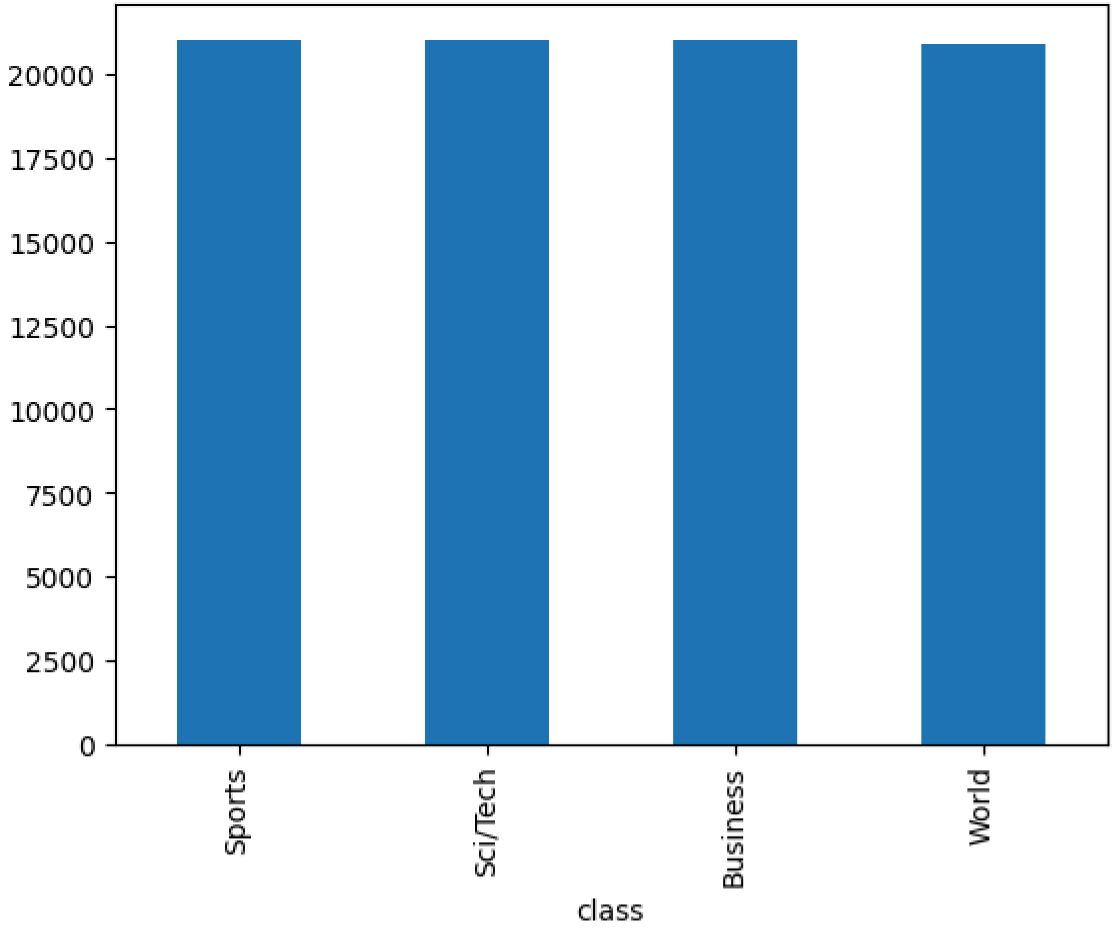
In [4]:

```
#Se grafican para poder observar su balance
pd.value_counts(train_df['class']).plot.bar()
```

```
/tmp/ipykernel_30/2157117126.py:2: FutureWarning: pandas.value_counts is depre
cated and 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]:
```
#Nos ayuda a observar que se tienen \ en el texto, lo cual no es bueno
print(train_df.loc[0, 'description'])
```

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

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

In [6]:
```
#Convierte a minúsculas y reemplaza / con espacios para limpiar el texto.
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 ...
...	...	...	...	...	...
53857	1	World	FDA Accused of Silencing Vioxx Warnings	WASHINGTON - The Food and Drug Administration ...	fda accused of silencing vioxx warnings washin...
111476	2	Sports	Buckeyes won #39;t play in NCAA or NIT tourneys	COLUMBUS, Ohio Ohio State has sanctioned its m...	buckeyes won #39;t play in ncaa or nit tourney...
6343	3	Business	Rate hikes by Fed work in two ways	If you #39;ve noticed that the price of everyt...	rate hikes by fed work in two ways if you #39;...
20736	4	Sci/Tech	NASA Administrator Offers Support for Kennedy ...	The following is a statement from NASA Adminis...	nasa administrator offers support for kennedy ...
34378	2	Sports	Twins make it 3 straight	The Minnesota Twins clinched on a bus in 1991....	twins make it 3 straight the minnesota twins c...

84000 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.

Se tokenizan las palabras individuales despues de haber sido previamente limpiadas, se hace para que el modelo pueda funcional

In [7]:

```python
#Tokeniza nuestras oraciones para su posterior analisis, creando una nueva colu
from nltk.tokenize import word_tokenize

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

```
  0%|          | 0/84000 [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...
...	...	...	...	...	...	...
53857	1	World	FDA Accused of Silencing Vioxx Warnings	WASHINGTON - The Food and Drug Administration ...	fda accused of silencing vioxx warnings washin...	[fda, accused, of, silencing, vioxx, warnings,...
111476	2	Sports	Buckeyes won #39;t play in NCAA or NIT tourneys	COLUMBUS, Ohio Ohio State has sanctioned its m...	buckeyes won #39;t play in ncaa or nit tourney...	[buckeyes, won, #, 39, ;, t, play, in, ncaa, o...
6343	3	Business	Rate hikes by Fed work in two ways	If you #39;ve noticed that the price of everyt...	rate hikes by fed work in two ways if you #39;...	[rate, hikes, by, fed, work, in, two, ways, if...
20736	4	Sci/Tech	NASA Administrator Offers Support for Kennedy ...	The following is a statement from NASA Adminis...	nasa administrator offers support for kennedy ...	[nasa, administrator, offers, support, for, ke...
34378	2	Sports	Twins make it 3 straight	The Minnesota Twins clinched on a bus in 1991....	twins make it 3 straight the minnesota twins c...	[twins, make, it, 3, straight, the, minnesota,...

84000 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.

Se genera un vocabulario para lograr delimitar la informacion importante, evitando asi que el modelo sobre aprenda demasiado debibo a palabras innecesarias.

In [8]:

```
#Solo si palabras que se repiten mas de diez veces se tomaran en cuenta como pa
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: 16,248
```

Se transforma el texto en una representacion numerica en el el texto puede entender

In [9]:

```
#Obtiene la cantidad de veces que cada palabra en el vocabulario tiene una cuen
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%|          | 0/84000 [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...	{2490: 1, 166: 1, 11: 1, 198: 1, 6548: 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...	{1921: 2, 0: 2, 731: 1, 5115: 1, 2822: 1, 740:...
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, (...	{7028: 2, 2: 1, 508: 1, 600: 1, 4: 1, 194: 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...	{2696: 1, 1: 4, 418: 2, 4: 3, 1047: 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...	{156: 2, 630: 1, 1503: 1, 21: 1, 2055: 2, 9: 1...
...	...	...	...	...	...	...	...
53857	1	World	FDA Accused of Silencing Vioxx Warnings	WASHINGTON - The Food and Drug Administration ...	fda accused of silencing vioxx warnings washin...	[fda, accused, of, silencing, vioxx, warnings,...	{2624: 1, 616: 1, 6: 3, 0: 3, 1640: 2, 2738: 1...
111476	2	Sports	Buckeyes won #39;t play in NCAA or NIT tourneys	COLUMBUS, Ohio Ohio State has sanctioned its m...	buckeyes won #39;t play in ncaa or nit tourney...	[buckeyes, won, #, 39, ;, t, play, in, ncaa, o...	{7246: 2, 241: 1, 12: 2, 13: 2, 8: 2, 149: 1, ...
6343	3	Business	Rate hikes by Fed work in two ways	If you #39;ve noticed that the price of everyt...	rate hikes by fed work in two ways if you #39;...	[rate, hikes, by, fed, work, in, two, ways, if...	{645: 1, 3946: 1, 27: 1, 1385: 1, 365: 1, 7: 1...
20736	4	Sci/Tech	NASA Administrator Offers Support for Kennedy ...	The following is a statement from NASA Adminis...	nasa administrator offers support for kennedy ...	[nasa, administrator, offers, support, for, ke...	{421: 2, 5276: 2, 846: 1, 420: 1, 11: 1, 3684:...

	class index	class	title	description	text	tokens	features
34378	2	Sports	Twins make it 3 straight	The Minnesota Twins clinched on a bus in 1991....	twins make it 3 straight the minnesota twins c...	[twins, make, it, 3, straight, the, minnesota,...	{1982: 2, 204: 1, 29: 1, 424: 1, 556: 1, 1: 1,...

84000 rows × 7 columns

Las funciones x_train y Y_train se hacen tensores, para que puedan ser compatibles con el modelo.

In [10]:

```python
#Convierte el diccionario en un vector para que el modelo sea compatible
def make_dense(feats):
    x = np.zeros(vocabulary_size)
    for k,v in feats.items():
        x[k] = v
    return x

# Aplica la función make_dense apila los resultados en una matriz 2D
X_train = np.stack(train_df['features'].progress_map(make_dense))

# Convierte la columna 'class index' en un array de NumPy
y_train = train_df['class index'].to_numpy() - 1

# Convierte los datos de entrenamiento 'X_train' en un tensor de PyTorch
X_train = torch.tensor(X_train, dtype=torch.float32)

# Convierte las etiquetas de clase 'y_train' en un tensor de PyTorch
y_train = torch.tensor(y_train)
```

```
  0%|          | 0/84000 [00:00<?, ?it/s]
```

La capa inicial toma el tamaño del vocabulario (n_feats) y las neuronas de las cantidades de clases (n_classes), a su vez, se realiza CrossEntropyLoss para la clasificacion multicalse y el optimizador SGD para actualizar los datos

In [11]:

```python
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)

# Inicia el modelo, la funcion de perdida, el optimizador y el cargador de dat
model = nn.Linear(n_feats, n_classes).to(device)
loss_func = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(), lr=lr)

# Entrena el modelo
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%|          | 0/84000 [00:00<?, ?it/s]
epoch 2:   0%|          | 0/84000 [00:00<?, ?it/s]
epoch 3:   0%|          | 0/84000 [00:00<?, ?it/s]
epoch 4:   0%|          | 0/84000 [00:00<?, ?it/s]
epoch 5:   0%|          | 0/84000 [00:00<?, ?it/s]
```

Next, we evaluate on the test dataset

In [12]:

```
# Repite todo lo anterior, pero con los datos de test para garantizar consisten
test_df = pd.read_csv('/kaggle/input/agnews-pytorch-simple-embed-classif-90/AG_
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%|          | 0/7600 [00:00<?, ?it/s]
  0%|          | 0/7600 [00:00<?, ?it/s]
  0%|          | 0/7600 [00:00<?, ?it/s]
```

In [13]:

```
from sklearn.metrics import classification_report

# Evalua el modelo
model.eval()

# No guarda las gradientes
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       0.92      0.86      0.89      1900
      Sports       0.91      0.97      0.94      1900
    Business       0.80      0.87      0.84      1900
    Sci/Tech       0.88      0.81      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
```