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

# Habilita tqdm en pandas
tqdm.pandas()

# Pones en True para poder usar la gpu (Si hay una disponible)
use_gpu = True

# Seleciona un device
device = torch.device('cuda' if use_gpu and torch.cuda.is_available() else 'cpu')
print(f'device: {device.type}')

# Semilla random
seed = 1234

# Seleciona una semilla random
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 [2]:

```
train_df =
pd.read_csv('/kaggle/input/agnews-pytorch-simple-embed-classif-90/AG_NEWS/train.csv',
header=None) # leer el dataset que se usara
train_df.columns = ['class index', 'title', 'description'] # Crear las columnas que se usaran
train_df = train_df.sample(frac = 0.7, random_state = 42) # Elejir una fraccion de los datos
train_df
```

Out[2]:

	class index	title	description
71787	3	BBC set for major shake-up, claims newspaper	London - The British Broadcasting Corporation,...

	class index	title	description
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...

	class index	title	description
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.

In [3]:

```
labels = open('/kaggle/input/classes/classes.txt').read().splitlines() # Crear lables para almacenar todos los nombres de las clases
classes = train_df['class index'].map(lambda i: labels[i-1]) # Crear clases para almacenar todos los nombres de las clases
train_df.insert(1, 'class', classes) # Insertar los nombres de las clases en el data frame
train_df
```

Out[3]:

	class index	class	title	description
71787	3	Business	BBC set for major shake-up, claims	London - The British Broadcasting

	class index	class	title	description
			newspaper	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...

	class index	class	title	description
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.

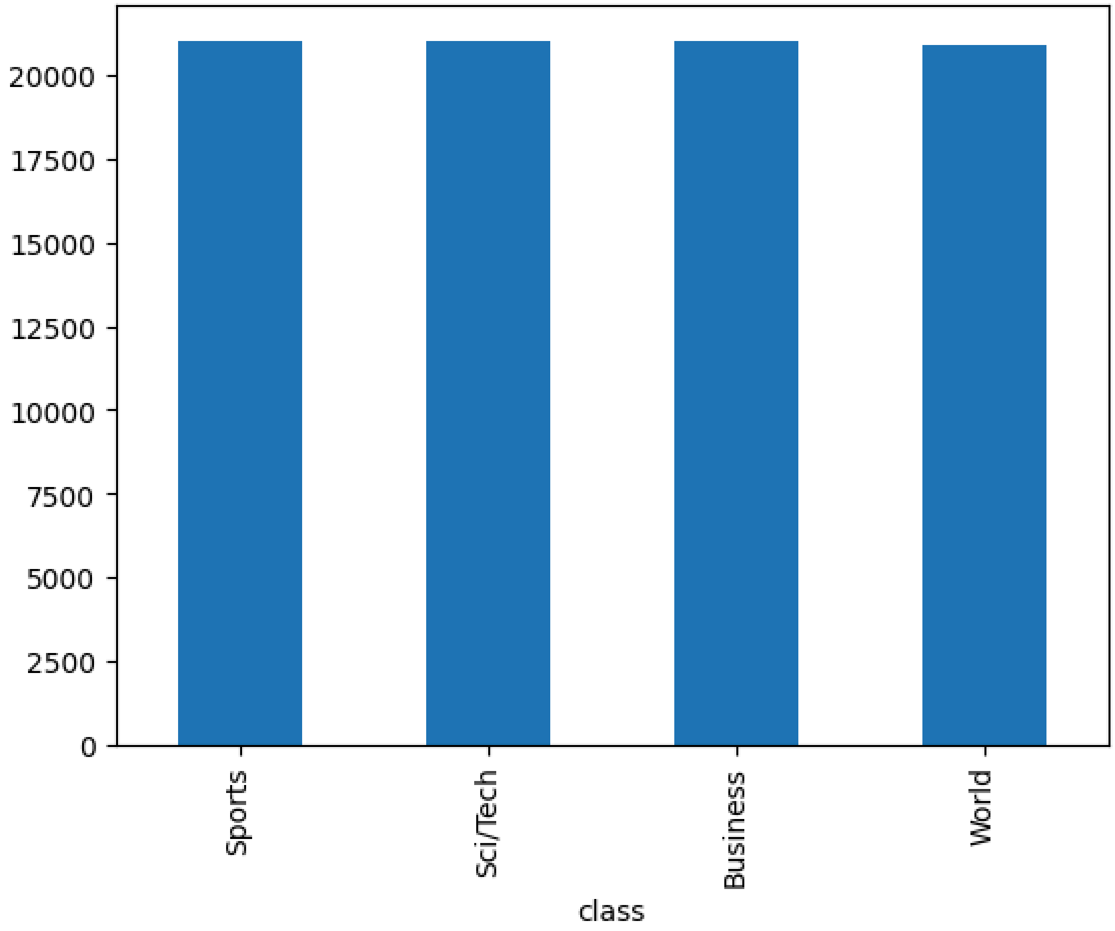
In [4]:

```
pd.value_counts(train_df['class']).plot.bar() # Se grafica pra ver como estan los resultados
```

```
/tmp/ipykernel_154/1245903889.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[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'])
```

```
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 [6]:

```
title = train_df['title'].str.lower() # Combertir los titulos en minusculas
descr = train_df['description'].str.lower() # Combertiri las descripciones en minusculas
text = title + " " + descr # Combainar title y descr en una columna de texto
train_df['text'] = text.str.replace('\\', ' ', regex=False) # Lipiar el texto con caracteres especificos
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...

	class index	class	title	description	text
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	COLUMBUS, Ohio	buckeyes won #39;t

	class index	class	title	description	text
			play in NCAA or NIT tourneys	Ohio State has sanctioned its m...	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.

In [7]:

```
from nltk.tokenize import word_tokenize

train_df['tokens'] = train_df['text'].progress_map(word_tokenize) # Tokenizacion de palabras y los almacena en una nueva columna
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	When the	flying the	[flying,

	class index	class	title	description	text	tokens
			Sun to Safety	Genesis capsule comes back to Earth w...	sun to safety when the genesis caps...	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...

	class index	class	title	description	text	tokens
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.

In [8]:

```
threshold = 10
tokens = train_df['tokens'].explode().value_counts() # Cuenta la frecuencia de cada token en la columna de tokens
tokens = tokens[tokens > threshold] # Muestra solo los tokens que superen la frecuencia que se termino al inicio
id_to_token = ['[UNK]'] + tokens.index.tolist() # Crea una lista de tokens unicos y se usa el UNK para las palabras poco frecuentes
token_to_id = {w:i for i,w in enumerate(id_to_token)} # Se crea un diccionario y se le asigna un ID unico a los tokens
```

```
vocabulary_size = len(id_to_token) # Obtienes la longitud de los id de los tokens
print(f'vocabulary size: {vocabulary_size:,}')
```

```
vocabulary size: 16,248
```

In [9]:

```
from collections import defaultdict

def make_feature_vector(tokens, unk_id=0): #  Crea una funcion en donde se crea un vector con las caracteristicas como si fuera un diccionario
    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,	London - The British Broadcasting	bbc set for major shake-up,	[bbc, set, for, major, shake-up, ,,	{2490: 1, 166: 1, 11: 1, 198: 1, 6548: 2,

	class index	class	title	description	text	tokens	features
			claims newspaper	Corporation,...	claims newspaper l...	claims, ne...	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	stocks seen flat as nortel and oil weigh	[stocks, seen, flat, as, nortel, and, oil, wei...	{156: 2, 630: 1, 1503: 1, 21: 1, 2055: 2, 9: 1...

	class index	class	title	description	text	tokens	features
				set to ...	new ...		
...	...	...	...	...	...	...	...
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	The following is a statement from	nasa administrator offers support	[nasa, administrator, offers, support,	{421: 2, 5276: 2, 846: 1, 420: 1, 11: 1,

	class index	class	title	description	text	tokens	features
			for Kennedy ...	NASA Adminis ...	for kennedy ...	for, ke...	3684:...
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

In [10]:

```
def make_dense(feats): # Se crea una funcion para poner los datos de train en PyTorch
    x = np.zeros(vocabulary_size)  # Crea un vector de ceros del tamaño del vocabulario.
    for k, v in feats.items():  # Itera sobre cada característica y su valor en el diccionario.
        x[k] = v
    return x

# Aplica la función make_dense a cada fila de la columna features del data frame de train_df y convierte el resultado en una matriz.
X_train = np.stack(train_df['features'].progress_map(make_dense))
y_train = train_df['class index'].to_numpy() - 1 # Convierte la columna class index en un array de NumPy, ajustando los índices para que inicien en 0.

# Convierte los datos de entrenamiento a tensors de PyTorch con el tipo de datos
X_train = torch.tensor(X_train, dtype=torch.float32)
y_train = torch.tensor(y_train)
```

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

In [11]:

```
from torch import nn
from torch import optim

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

# Inicializa el modelo con las funciones de 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)

# Entrenas 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}'):
        # Borra los gradientes acumulados
        model.zero_grad()
        # Envía el dato y la etiqueta al dispositivo adecuado
        x = X_train[i].unsqueeze(0).to(device)
        y_true = y_train[i].unsqueeze(0).to(device)
        # Predice las puntuaciones de etiquetas
        y_pred = model(x)
        # Calcula la pérdida (diferencia entre predicción y etiqueta real)
        loss = loss_func(y_pred, y_true)
        # Realiza la retropropagación
        loss.backward()
        # Optimiza los parámetros del modelo
        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 [13]:

```
# Repite todo el preproceso de arriba, pero ahora con el data set de test
test_df = 
pd.read_csv('/kaggle/input/agnews-pytorch-simple-embed-classif-90/AG_NEWS/test.csv', 
header=None)
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%|          | 0/7600 [00:00<?, ?it/s]
```

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

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

In [14]:

```
from sklearn.metrics import classification_report

# Se pone el modelo en modo evaluacion
model.eval()

# No se guardan los gradientes
with torch.no_grad():
    X_test = X_test.to(device) # Envía los datos al dispositivo (CPU o GPU)
    # Predice la clase más probable para cada ejemplo en el lote
    y_pred = torch.argmax(model(X_test), dim=1)
    # Convierte el tensor en un array numpy (y lo envía de regreso a la CPU si es necesario)
    y_pred = y_pred.cpu().numpy()
    # Imprime los resultados
    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
```