AG-News Text Classification using BERT:

In this project, we will cover in detail the application of BERT base model with respect to text classification. We will witness how this state of the art Transformer model is able to achieve extremely high performance metrics with respect to a large corpus of data comprising of more than 100k+ labelled training examples. The hugging face transformer & dataset library along with ktrain (a high level python wrapper with tensorflow backend) will be used to build, train & fine tune the BERT model with respect to classification on this custom dataset.

Checking Hardware Acceleration:

In [1]:

```
gpu_info = !nvidia-smi
gpu_info = '\n'.join(gpu_info)
if gpu_info.find('failed') >= 0:
  print('Select the Runtime > "Change runtime type" menu to enable a GPU accelerator, ')
  print('and then re-execute this cell.')
else:
  print(gpu_info)
```

```
Mon Jul 26 07:06:49 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 470.42.01    Driver Version: 460.32.03    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla P100-PCIE...  Off  | 00000000:00:04.0 Off |                    0 |
| N/A   46C    P0    30W / 250W |      0MiB / 16280MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

In [2]:

```
from psutil import virtual_memory
ram_gb = virtual_memory().total / 1e9
print('Your runtime has {:.1f} gigabytes of available RAM\n'.format(ram_gb))
```

```
Your runtime has 27.3 gigabytes of available RAM
```

Install Libraries:

```
In [3]: !pip install ktrain
        !pip install transformers
        !pip install datasets
```

```
Collecting ktrain
  Downloading ktrain-0.27.1.tar.gz (25.3 MB)
     |████████████████████████████████| 25.3 MB 58.4 MB/s
Collecting scikit-learn==0.23.2
  Downloading scikit_learn-0.23.2-cp37-cp37m-manylinux1_x86_64.whl (6.8 MB)
     |████████████████████████████████| 6.8 MB 47.8 MB/s
Requirement already satisfied: matplotlib>=3.0.0 in /usr/local/lib/python3.7/dist-packages (from ktrain) (3.2.2)
Requirement already satisfied: pandas>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from ktrain) (1.1.5)
Requirement already satisfied: fastprogress>=0.1.21 in /usr/local/lib/python3.7/dist-packages (from ktrain) (1.0.0)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from ktrain) (2.23.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from ktrain) (1.0.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from ktrain) (21.0)
Requirement already satisfied: ipython in /usr/local/lib/python3.7/dist-packages (from ktrain) (5.5.0)
Collecting langdetect
  Downloading langdetect-1.0.9.tar.gz (981 kB)
     |████████████████████████████████| 981 kB 39.1 MB/s
Requirement already satisfied: jieba in /usr/local/lib/python3.7/dist-packages (from ktrain) (0.42.1)
Collecting cchardet
  Downloading cchardet-2.1.7-cp37-cp37m-manylinux2010_x86_64.whl (263 kB)
     |████████████████████████████████| 263 kB 48.7 MB/s
```

```
In [5]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import ktrain
        from ktrain import text
        import tensorflow as tf
        from sklearn.model_selection import train_test_split
        from datasets import list_datasets
        from datasets import load_dataset
        import timeit
```

```
In [6]: print("Tensorflow version : ", tf.__version__)
        print("GPU available : ",bool(tf.test.is_gpu_available))
        print("GPU name : ",tf.test.gpu_device_name())
```

```
Tensorflow version :  2.5.0
GPU available :  True
GPU name :  /device:GPU:0
```

Checking available datasets in Hugging Face:

In [7]:
```
available_datasets = list_datasets()
print("Count of available datasets : ", len(available_datasets))
print()
print("<====== Dataset List ======> :\n")
print('\n  |__ '.join(dataset for dataset in available_datasets))
```

```
Count of available datasets :  1089

<====== Dataset List ======> :

acronym_identification
  |__ ade_corpus_v2
  |__ adversarial_qa
  |__ aeslc
  |__ afrikaans_ner_corpus
  |__ ag_news
  |__ ai2_arc
  |__ air_dialogue
  |__ ajgt_twitter_ar
  |__ allegro_reviews
  |__ allocine
  |__ alt
  |__ amazon_polarity
  |__ amazon_reviews_multi
  |__ amazon_us_reviews
```

Import AG News Dataset:

In [8]:

```
ag_news_dataset = load_dataset('ag_news')
print("\n", ag_news_dataset)
```

```
Downloading:   0%|          | 0.00/1.83k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/1.28k [00:00<?, ?B/s]

Using custom data configuration default

Downloading and preparing dataset ag_news/default (download: 29.88 MiB, generated: 30.23 MiB, post-processed: Unknown size, total: 60.10 MiB) to /root/.cache/huggingface/datas
ets/ag_news/default/0.0.0/bc2bcb40336ace1a0374767fc29bb0296cdaf8a6da7298436239c54d79180548...

Downloading:   0%|          | 0.00/11.0M [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/751k [00:00<?, ?B/s]

0 examples [00:00, ? examples/s]

0 examples [00:00, ? examples/s]

Dataset ag_news downloaded and prepared to /root/.cache/huggingface/datasets/ag_news/default/0.0.0/bc2bcb40336ace1a0374767fc29bb0296cdaf8a6da7298436239c54d79180548. Subsequent
calls will reuse this data.

 DatasetDict({
    train: Dataset({
        features: ['text', 'label'],
        num_rows: 120000
    })
    test: Dataset({
        features: ['text', 'label'],
        num_rows: 7600
    })
})
```

Dataset Details:

AG is a collection of more than 1 million news articles. News articles have been gathered from more than 2000 news sources by ComeToMyHead in more than 1 year of activity. ComeToMyHead is an academic news search engine which has been running since July, 2004. The dataset is provided by the academic comunity for research purposes in data mining (clustering, classification, etc), information retrieval (ranking, search, etc), xml, data compression, data streaming, and any other non-commercial activity. For more information, please refer to the link http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html (http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html)

AG News (AG’s News Corpus) is a subdataset of AG's corpus of news articles constructed by assembling titles and description fields of articles from the 4 largest classes (“World”, “Sports”, “Business”, “Sci/Tech”) of AG’s Corpus. The AG News contains 30,000 training and 1,900 test samples per class.

In [9]:

```
print("Dataset Items: \n", ag_news_dataset.items())
print("\nDataset type: \n", type(ag_news_dataset))
print("\nShape of dataset: \n", ag_news_dataset.shape)
print("\nNo of rows: \n", ag_news_dataset.num_rows)
print("\nNo of columns: \n", ag_news_dataset.num_columns)
```

```
Dataset Items: 
 dict_items([('train', Dataset({
    features: ['text', 'label'],
    num_rows: 120000
})), ('test', Dataset({
    features: ['text', 'label'],
    num_rows: 7600
}))])

Dataset type: 
 <class 'datasets.dataset_dict.DatasetDict'>

Shape of dataset: 
 {'train': (120000, 2), 'test': (7600, 2)}

No of rows: 
 {'train': 120000, 'test': 7600}

No of columns: 
 {'train': 2, 'test': 2}
```

In [10]:

```
print("\nColumn Names: \n", ag_news_dataset.column_names)
print("\n", ag_news_dataset.data)
```

```

Column Names: 
 {'train': ['text', 'label'], 'test': ['text', 'label']}

 {'train': MemoryMappedTable
text: string
label: int64, 'test': MemoryMappedTable
text: string
label: int64}
```

In [11]:

```
print(ag_news_dataset['train'][0])
print(ag_news_dataset['train'][1])
```

```
{'text': "Wall St. Bears Claw Back Into the Black (Reuters) Reuters - Short-sellers, Wall Street's dwindling\\band of ultra-cynics, are seeing green again.", 'label': 2}
{'text': 'Carlyle Looks Toward Commercial Aerospace (Reuters) Reuters - Private investment firm Carlyle Group,\\which has a reputation for making well-timed and occasionally
\\controversial plays in the defense industry, has quietly placed\\its bets on another part of the market.', 'label': 2}
```

In [12]:
```
print(ag_news_dataset['train']['text'][0])
print(ag_news_dataset['train']['label'][0])
print()
print(ag_news_dataset['train']['text'][35000])
print(ag_news_dataset['train']['label'][35000])
print()
print(ag_news_dataset['train']['text'][60000])
print(ag_news_dataset['train']['label'][60000])
print()
print(ag_news_dataset['train']['text'][100000])
print(ag_news_dataset['train']['label'][100000])
```

```
Wall St. Bears Claw Back Into the Black (Reuters) Reuters - Short-sellers, Wall Street's dwindling\band of ultra-cynics, are seeing green again.
2

Black armbands for Clough, tears for Liverpool fans In the afternoon, Brian Clough, unquestionably one of the greats and unarguably one of the most controversial of football m
en, died of cancer.
1

BYTE OF THE APPLE Apple lost one war to Microsoft by not licensing its Mac operating system. It may repeat the error with its iPod and music software.
3

Venezuelan Car-Bomb Suspect Killed, Weapons Found  CARACAS, Venezuela (Reuters) - A Venezuelan lawyer  suspected in last week's bombing murder of a top state  prosecutor was k
illed in a gunfight with police on Tuesday  after he tried to ram detectives with his car and opened fire  on them, officials said.
0
```

Loading Train & Test Datasets:

In [13]:
```
ag_news_train = load_dataset('ag_news', split='train')
ag_news_test = load_dataset('ag_news', split='test')
print("Train Dataset : ", ag_news_train.shape)
print("Test Dataset : ", ag_news_test.shape)
```

```
Using custom data configuration default
Reusing dataset ag_news (/root/.cache/huggingface/datasets/ag_news/default/0.0.0/bc2bcb40336ace1a0374767fc29bb0296cdaf8a6da7298436239c54d79180548)
Using custom data configuration default
Reusing dataset ag_news (/root/.cache/huggingface/datasets/ag_news/default/0.0.0/bc2bcb40336ace1a0374767fc29bb0296cdaf8a6da7298436239c54d79180548)

Train Dataset :  (120000, 2)
Test Dataset :  (7600, 2)
```

In [14]:
```
print(ag_news_train[0])
print(ag_news_test[0])
```

```
{'text': "Wall St. Bears Claw Back Into the Black (Reuters) Reuters - Short-sellers, Wall Street's dwindling\\band of ultra-cynics, are seeing green again.", 'label': 2}
{'text': "Fears for T N pension after talks Unions representing workers at Turner   Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul.",
'label': 2}
```

In [15]:
```
print("\nTrain Dataset Features: \n", ag_news_train.features)
print("\nTest Dataset Features: \n", ag_news_test.features)
```

```

Train Dataset Features:
 {'text': Value(dtype='string', id=None), 'label': ClassLabel(num_classes=4, names=['World', 'Sports', 'Business', 'Sci/Tech'], names_file=None, id=None)}

Test Dataset Features:
 {'text': Value(dtype='string', id=None), 'label': ClassLabel(num_classes=4, names=['World', 'Sports', 'Business', 'Sci/Tech'], names_file=None, id=None)}
```

Creating DataFrame object for K-train:

```
In [16]: pd.set_option('Display.max_columns', None)
         ag_news_train_df = pd.DataFrame(data=ag_news_train)
         ag_news_train_df.head(10)
```

```
Out[16]:
                                                  text  label
0          Wall St. Bears Claw Back Into the Black (Reute...      2
1     Carlyle Looks Toward Commercial Aerospace (Reu...      2
2       Oil and Economy Cloud Stocks' Outlook (Reuters...      2
3         Iraq Halts Oil Exports from Main Southern Pipe...      2
4             Oil prices soar to all-time record, posing new...      2
5      Stocks End Up, But Near Year Lows (Reuters) Re...      2
6       Money Funds Fell in Latest Week (AP) AP - Asse...      2
7        Fed minutes show dissent over inflation (USATO...      2
8         Safety Net (Forbes.com) Forbes.com - After ear...      2
9         Wall St. Bears Claw Back Into the Black NEW Y...      2
```

```
In [17]: ag_news_train_df.tail(10)
```

```
Out[17]:
                                                         text  label
119990          Barack Obama Gets #36;1.9 Million Book Deal (...      0
119991            Rauffer Beats Favorites to Win Downhill VAL G...      1
119992              Iraqis Face Winter Shivering by Candlelight B...      0
119993     AU Says Sudan Begins Troop Withdrawal from Dar...      0
119994       Syria Redeploys Some Security Forces in Lebano...      0
119995         Pakistan's Musharraf Says Won't Quit as Army C...      0
119996            Renteria signing a top-shelf deal Red Sox gene...      1
119997          Saban not going to Dolphins yet The Miami Dolp...      1
119998  Today's NFL games PITTSBURGH at NY GIANTS Time...      1
119999           Nets get Carter from Raptors INDIANAPOLIS -- A...      1
```

In [18]:

```
ag_news_test_df = pd.DataFrame(data=ag_news_test)
ag_news_test_df.head(10)
```

Out[18]:

	text	label
0	Fears for T N pension after talks Unions repre...	2
1	The Race is On: Second Private Team Sets Launc...	3
2	Ky. Company Wins Grant to Study Peptides (AP) ...	3
3	Prediction Unit Helps Forecast Wildfires (AP) ...	3
4	Calif. Aims to Limit Farm-Related Smog (AP) AP...	3
5	Open Letter Against British Copyright Indoctri...	3
6	Loosing the War on Terrorism \\"Sven Jaschan, ...	3
7	FOAFKey: FOAF, PGP, Key Distribution, and Bloo...	3
8	E-mail scam targets police chief Wiltshire Pol...	3
9	Card fraud unit nets 36,000 cards In its first...	3

In [19]:

```
ag_news_test_df.tail(10)
```

Out[19]:

	text	label
7590	Saban hiring on hold DAVIE - The Dolphins want...	1
7591	Bosnian-Serb prime minister resigns in protest...	0
7592	Historic Turkey-EU deal welcomed The European ...	0
7593	Mortaza strikes to lead superb Bangladesh rall...	1
7594	Powell pushes diplomacy for N. Korea WASHINGTO...	0
7595	Around the world Ukrainian presidential candid...	0
7596	Void is filled with Clement With the supply of...	1
7597	Martinez leaves bitter Like Roger Clemens did ...	1
7598	5 of arthritis patients in Singapore take Bext...	2
7599	EBay gets into rentals EBay plans to buy the a...	2

Data Preprocessing:

In [20]:

```
class_label_names = ['World', 'Sports', 'Business', 'Sci/Tech']
```

In [21]:

```
(X_train, y_train), (X_test, y_test), preprocessing_var = text.texts_from_df(train_df=ag_news_train_df,
                                                                             text_column='text',
                                                                             label_columns='label',
                                                                             val_df=ag_news_test_df,
                                                                             maxlen=512,
                                                                             preprocess_mode='bert')
```

```
['label_0', 'label_1', 'label_2', 'label_3']
   label_0  label_1  label_2  label_3
0      0.0      0.0      1.0      0.0
1      0.0      0.0      1.0      0.0
2      0.0      0.0      1.0      0.0
3      0.0      0.0      1.0      0.0
4      0.0      0.0      1.0      0.0
['label_0', 'label_1', 'label_2', 'label_3']
   label_0  label_1  label_2  label_3
0      0.0      0.0      1.0      0.0
1      0.0      0.0      0.0      1.0
2      0.0      0.0      0.0      1.0
3      0.0      0.0      0.0      1.0
4      0.0      0.0      0.0      1.0
downloading pretrained BERT model (uncased_L-12_H-768_A-12.zip)...
[██████████████████████████████████████████████████]
extracting pretrained BERT model...
done.

cleanup downloaded zip...
done.

preprocessing train...
language: en

done.

Is Multi-Label? False
preprocessing test...
language: en

done.
```

Create BERT Model:

In [22]:

```
transformer_bert_model = text.text_classifier(name='bert',
                                              train_data=(X_train, y_train),
                                              preproc=preprocessing_var)
```

```
Is Multi-Label? False
maxlen is 512
done.
```

```
In [23]: transformer_bert_model.layers
```

```
Out[23]: [<tensorflow.python.keras.engine.input_layer.InputLayer at 0x7f9352fc2390>,
          <tensorflow.python.keras.engine.input_layer.InputLayer at 0x7f93c51e2250>,
          <keras_bert.layers.embedding.TokenEmbedding at 0x7f9352fc2690>,
          <tensorflow.python.keras.layers.embeddings.Embedding at 0x7f935263c790>,
          <tensorflow.python.keras.layers.merge.Add at 0x7f9352eb0810>,
          <keras_pos_embd.pos_embd.PositionEmbedding at 0x7f9352dc2310>,
          <tensorflow.python.keras.layers.core.Dropout at 0x7f93532ae250>,
          <keras_layer_normalization.layer_normalization.LayerNormalization at 0x7f93526ef290>,
          <keras_multi_head.multi_head_attention.MultiHeadAttention at 0x7f9352735490>,
          <tensorflow.python.keras.layers.core.Dropout at 0x7f9352e73450>,
          <tensorflow.python.keras.layers.merge.Add at 0x7f9352f06390>,
          <keras_layer_normalization.layer_normalization.LayerNormalization at 0x7f9352e9fa50>,
          <keras_position_wise_feed_forward.feed_forward.FeedForward at 0x7f93526ef550>,
          <tensorflow.python.keras.layers.core.Dropout at 0x7f9352f00d90>,
          <tensorflow.python.keras.layers.merge.Add at 0x7f9352d55650>,
          <keras_layer_normalization.layer_normalization.LayerNormalization at 0x7f93c19c0fd0>,
          <keras_multi_head.multi_head_attention.MultiHeadAttention at 0x7f9352d11410>,
          <tensorflow.python.keras.layers.core.Dropout at 0x7f9352d5e410>,
          <tensorflow.python.keras.layers.merge.Add at 0x7f9352eeb2d0>,
```

Compile and train Bert in a Learner Object:

```
In [24]: bert_learner = ktrain.get_learner(model=transformer_bert_model,
                                           train_data=(X_train, y_train),
                                           val_data=(X_test, y_test),
                                           batch_size=6)
```

Best Hyper-parameters for BERT:

- Batch size: 16, 32
- Learning rate: 5e-5, 3e-5, 2e-5
- Number of epochs: 2, 3, 4

Train BERT on AG-News dataset:

```
In [27]: training_start_time = timeit.default_timer()
         bert_learner.fit_onecycle(lr=2e-5, epochs=3)
         training_stop_time = timeit.default_timer()
```

```
begin training using onecycle policy with max lr of 2e-05...
Epoch 1/3
20000/20000 [==============================] - 11620s 581ms/step - loss: 0.2264 - accuracy: 0.9233 - val_loss: 0.1768 - val_accuracy: 0.9412
Epoch 2/3
20000/20000 [==============================] - 11608s 580ms/step - loss: 0.1561 - accuracy: 0.9464 - val_loss: 0.1686 - val_accuracy: 0.9441
Epoch 3/3
20000/20000 [==============================] - 11601s 580ms/step - loss: 0.0832 - accuracy: 0.9714 - val_loss: 0.1582 - val_accuracy: 0.9500
```

```
In [29]: print("Total training time in minutes: \n", (training_stop_time - training_start_time)/60)
         print("Total training time in hours: \n", (training_stop_time - training_start_time)/3600)
```

```
Total training time in minutes:
 580.4977075469501
Total training time in hours:
 9.674961792449167
```

Checking BERT performance metrics:

```
In [30]: bert_learner.validate()
```

```
              precision    recall  f1-score   support

           0       0.97      0.96      0.96      1900
           1       0.99      0.99      0.99      1900
           2       0.93      0.92      0.92      1900
           3       0.92      0.94      0.93      1900

    accuracy                           0.95      7600
   macro avg       0.95      0.95      0.95      7600
weighted avg       0.95      0.95      0.95      7600
```

```
Out[30]: array([[1824,    6,   36,   34],
                [  10, 1875,    8,    7],
                [  32,    6, 1741,  121],
                [  23,    9,   88, 1780]])
```

```
In [31]: bert_learner.validate(class_names=class_label_names)
```

```
              precision    recall  f1-score   support

       World       0.97      0.96      0.96      1900
      Sports       0.99      0.99      0.99      1900
    Business       0.93      0.92      0.92      1900
    Sci/Tech       0.92      0.94      0.93      1900

    accuracy                           0.95      7600
   macro avg       0.95      0.95      0.95      7600
weighted avg       0.95      0.95      0.95      7600
```

```
Out[31]: array([[1824,    6,   36,   34],
                [  10, 1875,    8,    7],
                [  32,    6, 1741,  121],
                [  23,    9,   88, 1780]])
```

Saving the model:

```
In [32]: bert_predictor = ktrain.get_predictor(bert_learner.model, preproc=preprocessing_var)
         bert_predictor.get_classes()
```

```
Out[32]: ['label_0', 'label_1', 'label_2', 'label_3']
```

In [33]:

```
bert_predictor.save('/content/bert-ag-news-predictor')
```

```
/usr/local/lib/python3.7/dist-packages/tensorflow/python/keras/utils/generic_utils.py:497: CustomMaskWarning: Custom mask layers require a config and must override get_config.
When loading, the custom mask layer must be passed to the custom_objects argument.
  category=CustomMaskWarning)
```

In [34]:

```
!zip -r /content/bert-ag-news-predictor.zip /content/bert-ag-news-predictor
```

```
  adding: content/bert-ag-news-predictor/ (stored 0%)
  adding: content/bert-ag-news-predictor/tf_model.preproc (deflated 52%)
  adding: content/bert-ag-news-predictor/tf_model.h5 (deflated 11%)
```

Re-loading Model:

In [35]:

```
bert_predictor_2 = ktrain.load_predictor('/content/bert-ag-news-predictor')
bert_predictor_2.get_classes()
```

Out[35]:

```
['label_0', 'label_1', 'label_2', 'label_3']
```

References:

- https://huggingface.co/ (https://huggingface.co/)
- https://arxiv.org/abs/1810.04805 (https://arxiv.org/abs/1810.04805)

In [ ]: