Lab 6

Imports

```
!pip install transformers datasets simpletransformers==0.65.1
import pandas as pd
import numpy as np
import sklearn
from sklearn.metrics import classification_report
from simpletransformers.classification import ClassificationModel, Classification
import matplotlib.pyplot as plt
import seaborn as sn
from torch.utils.data import TensorDataset, DataLoader, RandomSampler, Sequential
import torch
from transformers import RobertaTokenizer, RobertaForSequenceClassification, Trai
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-
Collecting datasets
  Downloading datasets-2.18.0-py3-none-any.whl (510 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 510.5/510.5 kB 3.5 MB/s eta 0:00
Collecting simpletransformers==0.65.1
  Downloading simpletransformers-0.65.1-py3-none-any.whl (312 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 312.6/312.6 kB 3.3 MB/s eta 0:00
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: tqdm>=4.47.0 in /usr/local/lib/python3.10/dist-
Requirement already satisfied: regex in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-
Collecting seqeval (from simpletransformers==0.65.1)
  Downloading seqeval-1.2.2.tar.gz (43 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.6/43.6 kB 5.8 MB/s eta 0:00:0
  Preparing metadata (setup.py) ... done
Requirement already satisfied: tensorboard in /usr/local/lib/python3.10/dist-p
Collecting tensorboardx (from simpletransformers==0.65.1)
  Downloading tensorboardX-2.6.2.2-py2.py3-none-any.whl (101 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 101.7/101.7 kB 10.1 MB/s eta 0:0
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: tokenizers in /usr/local/lib/python3.10/dist-pa
Collecting wandb>=0.10.32 (from simpletransformers==0.65.1)
  Downloading wandb-0.16.4-py3-none-any.whl (2.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.2/2.2 MB 8.2 MB/s eta 0:00:00
Collecting streamlit (from simpletransformers==0.65.1)
  Downloading streamlit-1.32.2-py2.py3-none-any.whl (8.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.1/8.1 MB 5.2 MB/s eta 0:00:00
Requirement already satisfied: sentencepiece in /usr/local/lib/python3.10/dist
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: huggingface-hub<1.0,>=0.19.3 in /usr/local/lib/
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/di
```

```
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-p
Requirement already satisfied: safetensors>=0.4.1 in /usr/local/lib/python3.10
Requirement already satisfied: pyarrow>=12.0.0 in /usr/local/lib/python3.10/di
Requirement already satisfied: pyarrow-hotfix in /usr/local/lib/python3.10/dis
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl (116 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.3/116.3 kB 3.5 MB/s eta 0:00
Collecting xxhash (from datasets)
  Downloading xxhash-3.4.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.1/194.1 kB 3.1 MB/s eta 0:00
Collecting multiprocess (from datasets)
  Downloading multiprocess-0.70.16-py310-none-any.whl (134 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.8/134.8 kB 699.5 kB/s eta 0:
Requirement already satisfied: fsspec[http]<=2024.2.0,>=2023.1.0 in /usr/local
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/c
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.1
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dis
Requirement already satisfied: async-timeout<5.0,>=4.0 in /usr/local/lib/pytho
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/py
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/pyth
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10
```

Code

```
from sklearn.datasets import fetch_20newsgroups

categories = ['alt.atheism', 'comp.graphics', 'sci.med', 'sci.space']

newsgroups_train = fetch_20newsgroups(subset='train', remove=('headers', 'footers',
newsgroups_test = fetch_20newsgroups(subset='test', remove=('headers', 'footers', '
```

```
tokenizer = RobertaTokenizer.from_pretrained('roberta-base')
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:88: Us
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access pu
  warnings.warn(
```

tokenizer_config.json: 100% 25.0/25.0 [00:00<00:00, 1.02kB/s]

vocab.json: 100% 899k/899k [00:00<00:00, 11.8MB/s]

merges.txt: 100% 456k/456k [00:00<00:00, 20.7MB/s]

tokenizer.json: 100% 1.36M/1.36M [00:00<00:00, 42.3MB/

tokenizer.json: 100%

s]

```
train_df = pd.DataFrame({
    'text': newsgroups_train.data,
    'label': newsgroups_train.target
})

test_df = pd.DataFrame({
    'text': newsgroups_test.data,
    'label': newsgroups_test.target
})



model_args = ClassificationArgs()
model_args.num_train_epochs = 3
model_args.learning_rate = 4e-5
model_args.overwrite_output_dir = True
model_args.train_batch_size = 16
model_args.eval_batch_size = 8

model = ClassificationModel(
    "roberta",
    "roberta-base",
    num_labels=len(categories),
    args=model_args,
    use_cuda=True,
)
model.train_model(train_df)
```

model.safetensors: 100% 499M/499M [00:03<00:00, 133MB/s]

```
Some weights of RobertaForSequenceClassification were not initialized from the
You should probably TRAIN this model on a down-stream task to be able to use i
/usr/local/lib/python3.10/dist-packages/simpletransformers/classification/clas
  warnings.warn(
```

5/? [00:06<00:00, 1.22s/it]

Epoch 3 of 3: 100% 3/3 [01:26<00:00, 28.46s/it]

Epochs 1/3. Running Loss: 0.2361: 100% 141/141 [00:25<00:00, 7.05it/s]

Epochs 2/3. Running Loss: 0.0202: 100% 141/141 [00:21<00:00, 7.09it/s]

```
result, model_outputs, wrong_predictions = model.eval_model(test_df, verbose=True)
```

```
/usr/local/lib/python3.10/dist-packages/simpletransformers/classification/clas
  warnings.warn(
```

3/? [00:02<00:00, 1.29it/s]

Running Evaluation: 100% 188/188 [00:04<00:00, 45.81it/s]

```
predictions, raw_outputs = model.predict(test_df['text'].tolist())
print(classification_report(test_df['label'], predictions, target_names=categories
```

3/? [00:03<00:00, 1.05it/s]

100% 188/188 [00:06<00:00, 40.16it/s]

```
               precision    recall  f1-score   support

   alt.atheism       0.82      0.81      0.82       319
 comp.graphics       0.91      0.92      0.92       389
       sci.med       0.89      0.89      0.89       396
     sci.space       0.84      0.84      0.84       394

      accuracy                           0.87      1498
     macro avg       0.86      0.86      0.86      1498
  weighted avg       0.87      0.87      0.87      1498
```

Analysis

Summary of Model Performances

BERT (Lab6.4):

- **Precision**: Ranges from 0.83 to 0.90 across categories.
- **Recall**: Ranges from 0.79 to 0.89.
- **F1-Score**: Ranges from 0.81 to 0.90.
- **Overall Accuracy**: 0.85.

RoBERTa (Lab6_g47):

- **Precision**: Ranges from 0.78 to 0.92.
- **Recall**: Ranges from 0.81 to 0.89.
- **F1-Score**: Ranges from 0.79 to 0.91.
- **Overall Accuracy**: 0.85.

SVM (ConventionalSVM):

- **Precision**: Ranges from 0.74 to 0.88.
- **Recall**: Ranges from 0.76 to 0.87.

- **F1-Score**: Ranges from 0.80 to 0.87.
- **Overall Accuracy**: 0.83.

Analysis

Accuracy and F1-Score:

Transformer models (BERT and RoBERTa) show superior overall accuracy and F1-scores compared to the SVM model, indicating a better balance between precision and recall.

Precision and Recall Trade-offs:

- BERT exhibits slightly higher precision for 'alt.atheism' and 'sci.med', suggesting efficient identification of relevant instances.
- RoBERTa demonstrates higher recall in certain categories, indicating its effectiveness in retrieving more relevant instances, likely due to its advanced contextual understanding.
- SVM, while competitive, generally shows lower metrics, particularly in 'sci.space', possibly due to limitations in handling ambiguous content with bag-of-words features.

  Model Suitability:

- Transformer Models (BERT and RoBERTa): Best suited for tasks requiring deep textual understanding and contextualization, benefiting from their pre-training on extensive text corpora.
- SVM: Suitable for scenarios with limited computational resources or where model interpretability is crucial. It remains a robust baseline for simpler text classification problems.

Conclusion

Transformer-based models, BERT and RoBERTa, outperform the conventional SVM approach in the text classification task, highlighting their superior language understanding capabilities. The choice of model should, however, consider the specific requirements of the task, such as computational constraints, interpretability, and the complexity of the text data.