Group17Lab6

March 16, 2025

0.1 Lab6-Assignment: Topic Classification

Use the same training, development, and test partitions of the the 20 newsgroups text dataset as in Lab6.4-Topic-classification-BERT.ipynb

- Fine-tune and examine the performance of another transformer-based pretrained language models, e.g., RoBERTa, XLNet
- Compare the performance of this model to the results achieved in Lab6.4-Topic-classification-BERT.ipynb and to a conventional machine learning approach (e.g., SVM, Naive Bayes) using bag-of-words or other engineered features of your choice. Describe the differences in performance in terms of Precision, Recall, and F1-score evaluation metrics.

1 Transformer-based pretrained Language Model: RoBERTa

[22]:
```
pip install simpletransformers --upgrade
```

```
Requirement already satisfied: simpletransformers in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (0.70.1)
Requirement already satisfied: numpy in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(1.26.4)
Requirement already satisfied: requests in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(2.32.3)
Requirement already satisfied: tqdm>=4.47.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(4.67.1)
Requirement already satisfied: regex in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(2024.11.6)
Requirement already satisfied: transformers>=4.31.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(4.49.0)
Requirement already satisfied: datasets in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(3.3.2)
Requirement already satisfied: scipy in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
```

```
(1.13.1)
Requirement already satisfied: scikit-learn in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(1.6.1)
Requirement already satisfied: seqeval in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(1.2.2)
Requirement already satisfied: tensorboard in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(2.19.0)
Requirement already satisfied: tensorboardx in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(2.6.2.2)
Requirement already satisfied: pandas in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(2.2.3)
Requirement already satisfied: tokenizers in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(0.21.0)
Requirement already satisfied: wandb>=0.10.32 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(0.19.7)
Requirement already satisfied: streamlit in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(1.42.2)
Requirement already satisfied: sentencepiece in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from simpletransformers)
(0.2.0)
Requirement already satisfied: filelock in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
transformers>=4.31.0->simpletransformers) (3.13.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
transformers>=4.31.0->simpletransformers) (0.29.1)
Requirement already satisfied: packaging>=20.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
transformers>=4.31.0->simpletransformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
transformers>=4.31.0->simpletransformers) (6.0.2)
Requirement already satisfied: safetensors>=0.4.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
transformers>=4.31.0->simpletransformers) (0.5.3)
Requirement already satisfied: click!=8.0.0,>=7.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (8.1.7)
Requirement already satisfied: docker-pycreds>=0.4.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
```

```
wandb>=0.10.32->simpletransformers) (0.4.0)
Requirement already satisfied: gitpython!=3.1.29,>=1.0.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (3.1.44)
Requirement already satisfied: platformdirs in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (3.10.0)
Requirement already satisfied: protobuf!=4.21.0,!=5.28.0,<6,>=3.19.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (5.29.3)
Requirement already satisfied: psutil>=5.0.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (5.9.0)
Requirement already satisfied: pydantic<3,>=2.6 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (2.10.6)
Requirement already satisfied: sentry-sdk>=2.0.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (2.22.0)
Requirement already satisfied: setproctitle in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (1.3.5)
Requirement already satisfied: setuptools in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
wandb>=0.10.32->simpletransformers) (75.8.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
requests->simpletransformers) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
requests->simpletransformers) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
requests->simpletransformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
requests->simpletransformers) (2025.1.31)
Requirement already satisfied: pyarrow>=15.0.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
datasets->simpletransformers) (19.0.1)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
datasets->simpletransformers) (0.3.8)
Requirement already satisfied: xxhash in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
datasets->simpletransformers) (3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
```

```
datasets->simpletransformers) (0.70.16)
Requirement already satisfied: fsspec<=2024.12.0,>=2023.1.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
fsspec[http]<=2024.12.0,>=2023.1.0->datasets->simpletransformers) (2024.12.0)
Requirement already satisfied: aiohttp in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
datasets->simpletransformers) (3.11.13)
Requirement already satisfied: python-dateutil>=2.8.2 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
pandas->simpletransformers) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
pandas->simpletransformers) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
pandas->simpletransformers) (2025.1)
Requirement already satisfied: joblib>=1.2.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from scikit-
learn->simpletransformers) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from scikit-
learn->simpletransformers) (3.5.0)
Requirement already satisfied: altair<6,>=4.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (5.5.0)
Requirement already satisfied: blinker<2,>=1.0.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (1.9.0)
Requirement already satisfied: cachetools<6,>=4.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (5.5.2)
Requirement already satisfied: pillow<12,>=7.1.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (11.1.0)
Requirement already satisfied: rich<14,>=10.14.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (13.9.4)
Requirement already satisfied: tenacity<10,>=8.1.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (9.0.0)
Requirement already satisfied: toml<2,>=0.10.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (0.10.2)
Requirement already satisfied: typing-extensions<5,>=4.4.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (4.12.2)
Requirement already satisfied: pydeck<1,>=0.8.0b4 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
```

```
streamlit->simpletransformers) (0.9.1)
Requirement already satisfied: tornado<7,>=6.0.3 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
streamlit->simpletransformers) (6.4.2)
Requirement already satisfied: absl-py>=0.4 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
tensorboard->simpletransformers) (2.1.0)
Requirement already satisfied: grpcio>=1.48.2 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
tensorboard->simpletransformers) (1.70.0)
Requirement already satisfied: markdown>=2.6.8 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
tensorboard->simpletransformers) (3.7)
Requirement already satisfied: six>1.9 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
tensorboard->simpletransformers) (1.16.0)
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
tensorboard->simpletransformers) (0.7.2)
Requirement already satisfied: werkzeug>=1.0.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
tensorboard->simpletransformers) (3.1.3)
Requirement already satisfied: jinja2 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
altair<6,>=4.0->streamlit->simpletransformers) (3.1.5)
Requirement already satisfied: jsonschema>=3.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
altair<6,>=4.0->streamlit->simpletransformers) (4.23.0)
Requirement already satisfied: narwhals>=1.14.2 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
altair<6,>=4.0->streamlit->simpletransformers) (1.28.0)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
aiohttp->datasets->simpletransformers) (2.4.6)
Requirement already satisfied: aiosignal>=1.1.2 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
aiohttp->datasets->simpletransformers) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
aiohttp->datasets->simpletransformers) (24.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
aiohttp->datasets->simpletransformers) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
aiohttp->datasets->simpletransformers) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
```

```
aiohttp->datasets->simpletransformers) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
aiohttp->datasets->simpletransformers) (1.18.3)
Requirement already satisfied: gitdb<5,>=4.0.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
gitpython!=3.1.29,>=1.0.0->wandb>=0.10.32->simpletransformers) (4.0.12)
Requirement already satisfied: annotated-types>=0.6.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
pydantic<3,>=2.6->wandb>=0.10.32->simpletransformers) (0.7.0)
Requirement already satisfied: pydantic-core==2.27.2 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
pydantic<3,>=2.6->wandb>=0.10.32->simpletransformers) (2.27.2)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
rich<14,>=10.14.0->streamlit->simpletransformers) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
rich<14,>=10.14.0->streamlit->simpletransformers) (2.15.1)
Requirement already satisfied: MarkupSafe>=2.1.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
werkzeug>=1.0.1->tensorboard->simpletransformers) (3.0.2)
Requirement already satisfied: smmap<6,>=3.0.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
gitdb<5,>=4.0.1->gitpython!=3.1.29,>=1.0.0->wandb>=0.10.32->simpletransformers)
(5.0.2)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
jsonschema>=3.0->altair<6,>=4.0->streamlit->simpletransformers) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
jsonschema>=3.0->altair<6,>=4.0->streamlit->simpletransformers) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from
jsonschema>=3.0->altair<6,>=4.0->streamlit->simpletransformers) (0.22.3)
Requirement already satisfied: mdurl~=0.1 in
/opt/anaconda3/envs/myenv/lib/python3.12/site-packages (from markdown-it-
py>=2.2.0->rich<14,>=10.14.0->streamlit->simpletransformers) (0.1.2)
Note: you may need to restart the kernel to use updated packages.
```

[3]:
```
# Import libraries
import pandas as pd
import gensim
import numpy as np
import sklearn
from sklearn.metrics import classification_report
```

```
from simpletransformers.classification import ClassificationModel,␣
↪ClassificationArgs
import matplotlib.pyplot as plt
import seaborn as sn
import nltk
from nltk.stem import WordNetLemmatizer
from sklearn import svm
```

[6]:
```
from sklearn.datasets import fetch_20newsgroups

# load only a sub-selection of the categories (4 in our case)
categories = ['alt.atheism', 'comp.graphics', 'sci.med', 'sci.space']

# remove the headers, footers and quotes (to avoid overfitting)
newsgroups_train = fetch_20newsgroups(subset='train', remove=('headers',␣
↪'footers', 'quotes'), categories=categories, random_state=42)
newsgroups_test = fetch_20newsgroups(subset='test', remove=('headers',␣
↪'footers', 'quotes'), categories=categories, random_state=42)
```

[7]:
```
train = pd.DataFrame({'text': newsgroups_train.data, 'labels': newsgroups_train.
↪target})
```

[8]:
```
test = pd.DataFrame({'text': newsgroups_test.data, 'labels': newsgroups_test.
↪target})
```

[9]:
```
from sklearn.model_selection import train_test_split

train, dev = train_test_split(train, test_size=0.1, random_state=0,
                              stratify=train[['labels']])
```

[28]:
```
# Model configuration # https://simpletransformers.ai/docs/usage/
↪#configuring-a-simple-transformers-model
model_args = ClassificationArgs()

model_args.overwrite_output_dir=True # overwrite existing saved models in the␣
↪same directory
model_args.evaluate_during_training=True # to perform evaluation while training␣
↪the model
# (eval data should be passed to the training method)

model_args.num_train_epochs=10 # number of epochs
model_args.train_batch_size=32 # batch size
model_args.learning_rate=4e-6 # learning rate
model_args.max_seq_length=256 # maximum sequence length
# Note! Increasing max_seq_len may provide better performance, but training time␣
↪will increase.
# For educational purposes, we set max_seq_len to 256.
```

```
# Early stopping to combat overfitting: https://simpletransformers.ai/docs/
↪tips-and-tricks/#using-early-stopping
model_args.use_early_stopping=True
model_args.early_stopping_delta=0.01 # "The improvement over best_eval_loss␣
↪necessary to count as a better checkpoint"
model_args.early_stopping_metric='eval_loss'
model_args.early_stopping_metric_minimize=True
model_args.early_stopping_patience=2
model_args.evaluate_during_training_steps=32 # how often you want to run␣
↪validation in terms of training steps (or batches)
```

[29]:
```
model = ClassificationModel('roberta', 'roberta-base', num_labels=4,␣
↪args=model_args, use_cuda=False) # CUDA is enabled
```

```
Some weights of RobertaForSequenceClassification were not initialized from the
model checkpoint at roberta-base and are newly initialized:
['classifier.dense.bias', 'classifier.dense.weight', 'classifier.out_proj.bias',
'classifier.out_proj.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
```

[30]:
```
_, history = model.train_model(train, eval_df=dev)
```

```
  0%|          | 0/4 [00:00<?, ?it/s]huggingface/tokenizers: The current process
just got forked, after parallelism has already been used. Disabling parallelism
to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
```

```
false)
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
5it [00:05,  1.15s/it]
Epoch 1 of 10:   0%|          | 0/10 [00:00<?, ?it/s]
huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:04,  4.47s/it]

huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:04,  4.25s/it]
Epochs 1/10. Running Loss:    1.3845: 100%|| 64/64 [09:36<00:00,
9.01s/it]
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
1it [00:03,  3.40s/it]
Epoch 2 of 10:  10%|         | 1/10 [09:50<1:28:31, 590.15s/it]
huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:03,  3.88s/it]

huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
```

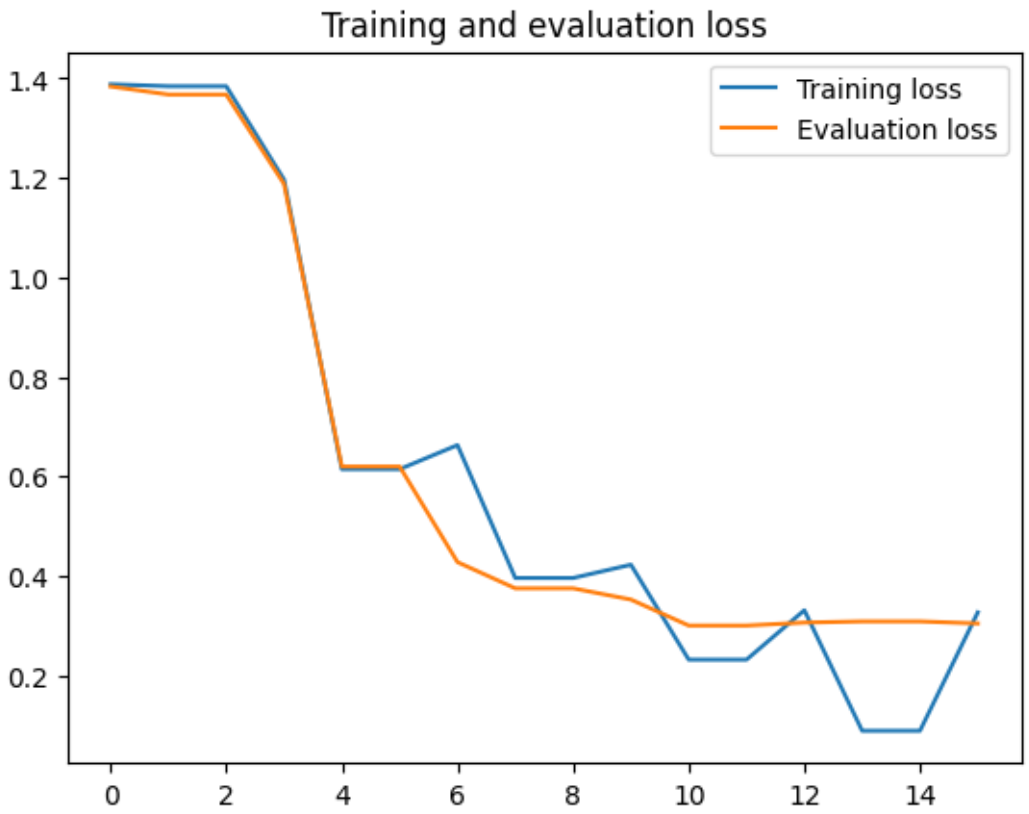
```
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:04,  4.14s/it]
Epochs 2/10. Running Loss:    0.6142: 100%|| 64/64 [09:32<00:00,
8.95s/it]
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
1it [00:03,  3.32s/it]
Epoch 3 of 10:  20%|        | 2/10 [19:36<1:18:23, 587.96s/it]
huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:03,  3.96s/it]

huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:03,  3.69s/it]
Epochs 3/10. Running Loss:    0.3953: 100%|| 64/64 [09:27<00:00,
8.86s/it]
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
1it [00:03,  3.10s/it]
Epoch 4 of 10:  30%|       | 3/10 [29:16<1:08:11, 584.46s/it]
huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
```

```
1it [00:03,  3.95s/it]

huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:03,  3.99s/it]
Epochs 4/10. Running Loss:    0.2310: 100%|| 64/64 [09:27<00:00,
8.87s/it]
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
1it [00:03,  3.12s/it]
Epoch 5 of 10:  40%|      | 4/10 [38:57<58:18, 583.06s/it]
huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:03,  3.96s/it]

huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:03,  3.95s/it]
Epochs 5/10. Running Loss:    0.0880: 100%|| 64/64 [09:28<00:00,
8.89s/it]
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
1it [00:03,  3.16s/it]
Epoch 6 of 10:  50%|     | 5/10 [48:39<48:33, 582.74s/it]
```

```
huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)

1it [00:03,  3.93s/it]
Epoch 6 of 10:  50%|     | 5/10 [53:28<53:28, 641.65s/it]
Epochs 6/10. Running Loss:    0.3262:  48%|     | 31/64 [04:48<05:06,
9.30s/it]
```

[37]:
```
# Training and evaluation loss
train_loss = history['train_loss']
eval_loss = history['eval_loss']
plt.plot(train_loss, label='Training loss')
plt.plot(eval_loss, label='Evaluation loss')
plt.title('Training and evaluation loss')
plt.legend()
```

[37]: <matplotlib.legend.Legend at 0x353093800>



[38]:
```
# Evaluate the model
result, model_outputs, wrong_predictions = model.eval_model(dev)
result
```

```
0it [00:00, ?it/s]huggingface/tokenizers: The current process just got forked,
after parallelism has already been used. Disabling parallelism to avoid
deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
1it [00:03,  3.53s/it]
Running Evaluation: 100%|| 3/3 [00:09<00:00,  3.23s/it]
```

[38]:
```
{'mcc': 0.8355903639998141, 'eval_loss': 0.30389564236005145}
```

[33]:
```
#Make predictions with the model (predict the labels of the documents in the␣
↪test set)
predicted, probabilities = model.predict(test.text.to_list())
test['predicted'] = predicted
```

```
  0%|          | 0/2 [00:00<?, ?it/s]huggingface/tokenizers: The current process
just got forked, after parallelism has already been used. Disabling parallelism
to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
huggingface/tokenizers: The current process just got forked, after parallelism
has already been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true |
false)
3it [00:04,  1.36s/it]
100%|| 15/15 [00:59<00:00,  3.96s/it]
```

[34]:
```
#test set predictions
test.head()
```

[34]:

	text	labels	predicted
0	\nAnd guess who's here in your place.\n\nPleas...	1	1

```
1  Does anyone know if any of Currier and Ives et...       1          1
2  =FLAME ON\n=\n=Reading through the posts about...       2          0
3  \nBut in this case I said I hoped that BCCI wa...       0          0
4  \nIn the kind I have made I used a Lite sour c...       2          2
```

[35]:
```
# Result (note: your result can be different due to randomness in operations)
print(classification_report(test['labels'], test['predicted']))
```

```
              precision    recall  f1-score   support

           0       0.83      0.82      0.82       319
           1       0.81      0.94      0.87       389
           2       0.91      0.87      0.89       396
           3       0.88      0.79      0.83       394

    accuracy                           0.86      1498
   macro avg       0.86      0.85      0.85      1498
weighted avg       0.86      0.86      0.86      1498
```

2 Conventional Machine Learning Approach: SVM

[10]:
```
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.svm import LinearSVC

# Text Preprocessing
lemmatizer = WordNetLemmatizer()
def lemmatize_stemming(text):
    return lemmatizer.lemmatize(text)

def preprocess(text):
    result = []
    for token in gensim.utils.simple_preprocess(text):
        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) >␣
 ↪3:
            result.append(lemmatize_stemming(token))
    return " ".join(result)  # Convert list of tokens back to string

# Apply preprocessing
train["processed_text"] = train["text"].map(preprocess)
test["processed_text"] = test["text"].map(preprocess)

# Convert text to BoW features using CountVectorizer
vectorizer = CountVectorizer()
X_train = vectorizer.fit_transform(train["processed_text"])
X_test = vectorizer.transform(test["processed_text"])
```

```python
# Get labels
y_train = train["labels"]
y_test = test["labels"]

# Train SVM classifier
lin_clf = LinearSVC()
lin_clf.fit(X_train, y_train)

# Evaluate on test set
y_test_pred = lin_clf.predict(X_test)
print(classification_report(y_test, y_test_pred))
```

```
              precision    recall  f1-score   support

           0       0.75      0.66      0.70       319
           1       0.72      0.83      0.77       389
           2       0.80      0.72      0.75       396
           3       0.70      0.74      0.72       394

    accuracy                           0.74      1498
   macro avg       0.74      0.73      0.74      1498
weighted avg       0.74      0.74      0.74      1498
```

```
/opt/anaconda3/envs/myenv/lib/python3.12/site-
packages/sklearn/svm/_base.py:1249: ConvergenceWarning: Liblinear failed to
converge, increase the number of iterations.
  warnings.warn(
```

3 Comparison

3.1 Classification Report: BERT Model

	precision	recall	f1-score	support
0	0.83	0.82	0.83	319
1	0.89	0.92	0.91	389
2	0.94	0.88	0.91	396
3	0.83	0.86	0.84	394
accuracy			0.87	1498
macro avg	0.87	0.87	0.87	1498
weighted avg	0.88	0.87	0.87	1498

3.2 Classification Report: RoBERTa Model

	precision	recall	f1-score	support
0	0.83	0.82	0.82	319
1	0.81	0.94	0.87	389
2	0.91	0.87	0.89	396
3	0.88	0.79	0.83	394
accuracy			0.86	1498
macro avg	0.86	0.85	0.85	1498
weighted avg	0.86	0.86	0.86	1498

3.3 Classification Report: SVM with BoW

	precision	recall	f1-score	support
0	0.75	0.66	0.70	319
1	0.72	0.83	0.77	389
2	0.80	0.72	0.75	396
3	0.70	0.74	0.72	394
accuracy			0.74	1498
macro avg	0.74	0.73	0.74	1498
weighted avg	0.74	0.74	0.74	1498

3.4 Comparison between the 3 models

Overall, the models BERT and RoBERTa performed better than the model that used SVM with BoW, with BERT slightly outperforming RoBERTa. For label 0, BERT and RoBERTa performed almost identically with an F1 score of 0.83 and 0.82 respectively, while SVM had an F1 score of 0.7. SVM also had the lowest recall (0.66) for label 0 which shows it failed to correctly identify actual instances of label 0.

For label 1, BERT performed better in terms of precision (0.89) and F1-score (0.91), whereas RoBERTa had a better recall (0.94). This shows that BERT made less false positive errors, while the recall of RoBERTa shows that it correctly identified many instances, but may have more false positives. On the other hand, the SVM model did have a fairly high recall (0.83), but compared to BERT and RoBERTa, it still did not perform as well.

For label 2, BERT had a better precision (0.94), recall (0.88) and F1-score (0.91) compared to RoBERTa and the SVM model.

For label 3, RoBERTa has a better precision (0.88), but a lower recall (0.79) which indicates that it has fewer false positives, but missed some true postives. In contrast, BERT identified more true postives of label 3, but may have made more false predictions. In the case of SVM, the precision (0.70) for label 3 is the lowest, which shows that SVM struggled the most with false positives for label 3.