# Sentiment Analysis

March 11, 2023

## 0.1 Bi-directional GRU Classifier with fastText embeddings using ktrain package

**Original Notebooks:** https://github.com/amaiya/ktrain

[2]:
```
!pip install tensorflow
```

```
Collecting tensorflow
  Downloading
tensorflow-2.11.0-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(588.3 MB)
                           588.3/588.3
MB 1.3 MB/s eta 0:00:0000:0100:01
Collecting tensorboard<2.12,>=2.11
  Downloading tensorboard-2.11.2-py3-none-any.whl (6.0 MB)
                           6.0/6.0 MB
21.2 MB/s eta 0:00:00:00:0100:01
Requirement already satisfied: protobuf<3.20,>=3.9.2 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (3.19.6)
Collecting h5py>=2.9.0
  Downloading
h5py-3.8.0-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.3 MB)
                           4.3/4.3 MB
9.2 MB/s eta 0:00:000:00:01
Collecting google-pasta>=0.1.1
  Downloading google_pasta-0.2.0-py3-none-any.whl (57 kB)
                           57.5/57.5 kB
1.7 MB/s eta 0:00:00
Requirement already satisfied: packaging in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (23.0)
Collecting keras<2.12,>=2.11.0
  Downloading keras-2.11.0-py2.py3-none-any.whl (1.7 MB)
                           1.7/1.7 MB
15.6 MB/s eta 0:00:00:00:01
Collecting wrapt>=1.11.0
  Downloading wrapt-1.15.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.man
ylinux_2_17_x86_64.manylinux2014_x86_64.whl (75 kB)
                           75.7/75.7 kB
```

```
270.6 kB/s eta 0:00:00a 0:00:01
Collecting astunparse>=1.6.0
  Downloading astunparse-1.6.3-py2.py3-none-any.whl (12 kB)
Requirement already satisfied: numpy>=1.20 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (1.21.6)
Requirement already satisfied: grpcio<2.0,>=1.24.3 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (1.51.1)
Collecting opt-einsum>=2.3.2
  Downloading opt_einsum-3.3.0-py3-none-any.whl (65 kB)
                             65.5/65.5 kB
2.6 MB/s eta 0:00:00
Collecting tensorflow-estimator<2.12,>=2.11.0
  Downloading tensorflow_estimator-2.11.0-py2.py3-none-any.whl (439 kB)
                            439.2/439.2
kB 1.8 MB/s eta 0:00:0000:01
Collecting tensorflow-io-gcs-filesystem>=0.23.1
  Downloading tensorflow_io_gcs_filesystem-0.31.0-cp37-cp37m-manylinux_2_12_x86_
64.manylinux2010_x86_64.whl (2.4 MB)
                             2.4/2.4 MB
5.3 MB/s eta 0:00:000:00:01
Collecting termcolor>=1.1.0
  Downloading termcolor-2.2.0-py3-none-any.whl (6.6 kB)
Requirement already satisfied: setuptools in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (66.1.1)
Requirement already satisfied: absl-py>=1.0.0 in /opt/conda/lib/python3.7/site-
packages (from tensorflow) (1.4.0)
Collecting gast<=0.4.0,>=0.2.1
  Downloading gast-0.4.0-py3-none-any.whl (9.8 kB)
Collecting flatbuffers>=2.0
  Downloading flatbuffers-23.3.3-py2.py3-none-any.whl (26 kB)
Requirement already satisfied: typing-extensions>=3.6.6 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (4.4.0)
Collecting libclang>=13.0.0
  Downloading libclang-15.0.6.1-py2.py3-none-manylinux2010_x86_64.whl (21.5 MB)
                             21.5/21.5 MB
65.8 MB/s eta 0:00:0000:0100:01
Requirement already satisfied: six>=1.12.0 in
/opt/conda/lib/python3.7/site-packages (from tensorflow) (1.16.0)
Requirement already satisfied: wheel<1.0,>=0.23.0 in
/opt/conda/lib/python3.7/site-packages (from astunparse>=1.6.0->tensorflow)
(0.38.4)
Requirement already satisfied: google-auth<3,>=1.6.3 in
/opt/conda/lib/python3.7/site-packages (from
tensorboard<2.12,>=2.11->tensorflow) (2.16.0)
Collecting google-auth-oauthlib<0.5,>=0.4.1
  Downloading google_auth_oauthlib-0.4.6-py2.py3-none-any.whl (18 kB)
Requirement already satisfied: requests<3,>=2.21.0 in
```

```
/opt/conda/lib/python3.7/site-packages (from
tensorboard<2.12,>=2.11->tensorflow) (2.28.2)
Collecting tensorboard-plugin-wit>=1.6.0
  Downloading tensorboard_plugin_wit-1.8.1-py3-none-any.whl (781 kB)
                             781.3/781.3 kB
54.2 MB/s eta 0:00:00
Collecting tensorboard-data-server<0.7.0,>=0.6.0
  Downloading tensorboard_data_server-0.6.1-py3-none-manylinux2010_x86_64.whl
(4.9 MB)
                              4.9/4.9 MB
91.9 MB/s eta 0:00:00:00:01
Collecting werkzeug>=1.0.1
  Downloading Werkzeug-2.2.3-py3-none-any.whl (233 kB)
                             233.6/233.6 kB
27.4 MB/s eta 0:00:00
Collecting markdown>=2.6.8
  Downloading Markdown-3.4.1-py3-none-any.whl (93 kB)
                              93.3/93.3 kB
13.3 MB/s eta 0:00:00
Requirement already satisfied: pyasn1-modules>=0.2.1 in
/opt/conda/lib/python3.7/site-packages (from google-
auth<3,>=1.6.3->tensorboard<2.12,>=2.11->tensorflow) (0.2.8)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in
/opt/conda/lib/python3.7/site-packages (from google-
auth<3,>=1.6.3->tensorboard<2.12,>=2.11->tensorflow) (5.3.0)
Requirement already satisfied: rsa<5,>=3.1.4 in /opt/conda/lib/python3.7/site-
packages (from google-auth<3,>=1.6.3->tensorboard<2.12,>=2.11->tensorflow) (4.9)
Requirement already satisfied: requests-oauthlib>=0.7.0 in
/opt/conda/lib/python3.7/site-packages (from google-auth-
oauthlib<0.5,>=0.4.1->tensorboard<2.12,>=2.11->tensorflow) (1.3.1)
Requirement already satisfied: importlib-metadata>=4.4 in
/opt/conda/lib/python3.7/site-packages (from
markdown>=2.6.8->tensorboard<2.12,>=2.11->tensorflow) (6.0.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/conda/lib/python3.7/site-packages (from
requests<3,>=2.21.0->tensorboard<2.12,>=2.11->tensorflow) (2.1.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/opt/conda/lib/python3.7/site-packages (from
requests<3,>=2.21.0->tensorboard<2.12,>=2.11->tensorflow) (1.26.14)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.7/site-
packages (from requests<3,>=2.21.0->tensorboard<2.12,>=2.11->tensorflow) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.7/site-packages (from
requests<3,>=2.21.0->tensorboard<2.12,>=2.11->tensorflow) (2022.12.7)
Requirement already satisfied: MarkupSafe>=2.1.1 in
/opt/conda/lib/python3.7/site-packages (from
werkzeug>=1.0.1->tensorboard<2.12,>=2.11->tensorflow) (2.1.2)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.7/site-
```

```
packages (from importlib-
metadata>=4.4->markdown>=2.6.8->tensorboard<2.12,>=2.11->tensorflow) (3.11.0)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in
/opt/conda/lib/python3.7/site-packages (from pyasn1-modules>=0.2.1->google-
auth<3,>=1.6.3->tensorboard<2.12,>=2.11->tensorflow) (0.4.8)
Requirement already satisfied: oauthlib>=3.0.0 in /opt/conda/lib/python3.7/site-
packages (from requests-oauthlib>=0.7.0->google-auth-
oauthlib<0.5,>=0.4.1->tensorboard<2.12,>=2.11->tensorflow) (3.2.2)
Installing collected packages: tensorboard-plugin-wit, libclang, flatbuffers,
wrapt, werkzeug, termcolor, tensorflow-io-gcs-filesystem, tensorflow-estimator,
tensorboard-data-server, opt-einsum, keras, h5py, google-pasta, gast,
astunparse, markdown, google-auth-oauthlib, tensorboard, tensorflow
  Attempting uninstall: google-auth-oauthlib
    Found existing installation: google-auth-oauthlib 0.8.0
    Uninstalling google-auth-oauthlib-0.8.0:
      Successfully uninstalled google-auth-oauthlib-0.8.0
Successfully installed astunparse-1.6.3 flatbuffers-23.3.3 gast-0.4.0 google-
auth-oauthlib-0.4.6 google-pasta-0.2.0 h5py-3.8.0 keras-2.11.0 libclang-15.0.6.1
markdown-3.4.1 opt-einsum-3.3.0 tensorboard-2.11.2 tensorboard-data-server-0.6.1
tensorboard-plugin-wit-1.8.1 tensorflow-2.11.0 tensorflow-estimator-2.11.0
tensorflow-io-gcs-filesystem-0.31.0 termcolor-2.2.0 werkzeug-2.2.3 wrapt-1.15.0
```

[2]:
```
import warnings
# warnings.filterwarnings('ignore')

import pandas as pd
import numpy as np
import re

from sklearn import preprocessing
from sklearn.model_selection import train_test_split

import tensorflow as tf
from keras.preprocessing.text import Tokenizer, text_to_word_sequence
```

```
2023-03-10 03:54:30.646172: I tensorflow/core/platform/cpu_feature_guard.cc:193]
This TensorFlow binary is optimized with oneAPI Deep Neural Network Library
(oneDNN) to use the following CPU instructions in performance-critical
operations:  AVX2 FMA
To enable them in other operations, rebuild TensorFlow with the appropriate
compiler flags.
2023-03-10 03:54:34.270416: W
tensorflow/compiler/xla/stream_executor/platform/default/dso_loader.cc:64] Could
not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.so.11.0: cannot
open shared object file: No such file or directory
2023-03-10 03:54:34.270477: I
tensorflow/compiler/xla/stream_executor/cuda/cudart_stub.cc:29] Ignore above
cudart dlerror if you do not have a GPU set up on your machine.
```

```
2023-03-10 03:54:41.691387: W
tensorflow/compiler/xla/stream_executor/platform/default/dso_loader.cc:64] Could
not load dynamic library 'libnvinfer.so.7'; dlerror: libnvinfer.so.7: cannot
open shared object file: No such file or directory
2023-03-10 03:54:41.692609: W
tensorflow/compiler/xla/stream_executor/platform/default/dso_loader.cc:64] Could
not load dynamic library 'libnvinfer_plugin.so.7'; dlerror:
libnvinfer_plugin.so.7: cannot open shared object file: No such file or
directory
2023-03-10 03:54:41.692629: W
tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Cannot
dlopen some TensorRT libraries. If you would like to use Nvidia GPU with
TensorRT, please make sure the missing libraries mentioned above are installed
properly.
```

### Check for GPU presence

[3]:
```
#Verify we got CPU + GPU or only CPU
tf.config.list_physical_devices()
```

```
2023-03-10 03:54:46.784434: W
tensorflow/compiler/xla/stream_executor/platform/default/dso_loader.cc:64] Could
not load dynamic library 'libcuda.so.1'; dlerror: libcuda.so.1: cannot open
shared object file: No such file or directory
2023-03-10 03:54:46.816208: W
tensorflow/compiler/xla/stream_executor/cuda/cuda_driver.cc:265] failed call to
cuInit: UNKNOWN ERROR (303)
2023-03-10 03:54:46.816290: I
tensorflow/compiler/xla/stream_executor/cuda/cuda_diagnostics.cc:156] kernel
driver does not appear to be running on this host (python-20230307-192621):
/proc/driver/nvidia/version does not exist
```

[3]:
```
[PhysicalDevice(name='/physical_device:CPU:0', device_type='CPU')]
```

[6]:
```
!nvidia-smi
```

```
/bin/bash: nvidia-smi: command not found
```

[8]:
```
!pip install ktrain
```

```
Collecting ktrain
  Downloading ktrain-0.33.2.tar.gz (25.3 MB)
                              25.3/25.3 MB
42.2 MB/s eta 0:00:0000:0100:01
  Preparing metadata (setup.py) … done
Requirement already satisfied: scikit-learn in
/opt/conda/lib/python3.7/site-packages (from ktrain) (1.0.2)
Requirement already satisfied: matplotlib>=3.0.0 in
/opt/conda/lib/python3.7/site-packages (from ktrain) (3.5.3)
```

```
Requirement already satisfied: pandas>=1.0.1 in /opt/conda/lib/python3.7/site-
packages (from ktrain) (1.3.5)
Collecting fastprogress>=0.1.21
  Downloading fastprogress-1.0.3-py3-none-any.whl (12 kB)
Requirement already satisfied: requests in /opt/conda/lib/python3.7/site-
packages (from ktrain) (2.28.2)
Requirement already satisfied: joblib in /opt/conda/lib/python3.7/site-packages
(from ktrain) (1.2.0)
Requirement already satisfied: packaging in /opt/conda/lib/python3.7/site-
packages (from ktrain) (23.0)
Collecting langdetect
  Downloading langdetect-1.0.9.tar.gz (981 kB)
                           981.5/981.5 kB
52.2 MB/s eta 0:00:00
  Preparing metadata (setup.py) … done
Collecting jieba
  Downloading jieba-0.42.1.tar.gz (19.2 MB)
                            19.2/19.2 MB
73.2 MB/s eta 0:00:0000:0100:01
  Preparing metadata (setup.py) … done
Collecting cchardet
  Downloading cchardet-2.1.7-cp37-cp37m-manylinux2010_x86_64.whl (263 kB)
                           263.7/263.7 kB
30.5 MB/s eta 0:00:00
Collecting chardet
  Downloading chardet-5.1.0-py3-none-any.whl (199 kB)
                           199.1/199.1 kB
23.2 MB/s eta 0:00:00
Collecting syntok>1.3.3
  Downloading syntok-1.4.4-py3-none-any.whl (24 kB)
Collecting transformers>=4.17.0
  Downloading transformers-4.26.1-py3-none-any.whl (6.3 MB)
                            6.3/6.3 MB
78.7 MB/s eta 0:00:00:00:0100:01
Collecting sentencepiece
  Downloading
sentencepiece-0.1.97-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(1.3 MB)
                            1.3/1.3 MB
59.6 MB/s eta 0:00:00
Collecting keras_bert>=0.86.0
  Downloading keras-bert-0.89.0.tar.gz (25 kB)
  Preparing metadata (setup.py) … done
Collecting whoosh
  Downloading Whoosh-2.7.4-py2.py3-none-any.whl (468 kB)
                           468.8/468.8 kB
41.3 MB/s eta 0:00:00
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-
```

```
packages (from keras_bert>=0.86.0->ktrain) (1.21.6)
Collecting keras-transformer==0.40.0
  Downloading keras-transformer-0.40.0.tar.gz (9.7 kB)
  Preparing metadata (setup.py) … done
Collecting keras-pos-embd==0.13.0
  Downloading keras-pos-embd-0.13.0.tar.gz (5.6 kB)
  Preparing metadata (setup.py) … done
Collecting keras-multi-head==0.29.0
  Downloading keras-multi-head-0.29.0.tar.gz (13 kB)
  Preparing metadata (setup.py) … done
Collecting keras-layer-normalization==0.16.0
  Downloading keras-layer-normalization-0.16.0.tar.gz (3.9 kB)
  Preparing metadata (setup.py) … done
Collecting keras-position-wise-feed-forward==0.8.0
  Downloading keras-position-wise-feed-forward-0.8.0.tar.gz (4.1 kB)
  Preparing metadata (setup.py) … done
Collecting keras-embed-sim==0.10.0
  Downloading keras-embed-sim-0.10.0.tar.gz (3.6 kB)
  Preparing metadata (setup.py) … done
Collecting keras-self-attention==0.51.0
  Downloading keras-self-attention-0.51.0.tar.gz (11 kB)
  Preparing metadata (setup.py) … done
Requirement already satisfied: pyparsing>=2.2.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->ktrain) (3.0.9)
Requirement already satisfied: kiwisolver>=1.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->ktrain) (1.4.4)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.7/site-
packages (from matplotlib>=3.0.0->ktrain) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->ktrain) (4.38.0)
Requirement already satisfied: python-dateutil>=2.7 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->ktrain) (2.8.2)
Requirement already satisfied: pillow>=6.2.0 in /opt/conda/lib/python3.7/site-
packages (from matplotlib>=3.0.0->ktrain) (9.4.0)
Requirement already satisfied: pytz>=2017.3 in /opt/conda/lib/python3.7/site-
packages (from pandas>=1.0.1->ktrain) (2022.7.1)
Requirement already satisfied: regex>2016 in /opt/conda/lib/python3.7/site-
packages (from syntok>1.3.3->ktrain) (2022.10.31)
Requirement already satisfied: pyyaml>=5.1 in /opt/conda/lib/python3.7/site-
packages (from transformers>=4.17.0->ktrain) (6.0)
Requirement already satisfied: filelock in /opt/conda/lib/python3.7/site-
packages (from transformers>=4.17.0->ktrain) (3.9.0)
Collecting tokenizers!=0.11.3,<0.14,>=0.11.1
  Downloading
tokenizers-0.13.2-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (7.6
MB)
                                    7.6/7.6 MB
94.5 MB/s eta 0:00:00:00:0100:01
```

```
Requirement already satisfied: importlib-metadata in
/opt/conda/lib/python3.7/site-packages (from transformers>=4.17.0->ktrain)
(6.0.0)
Collecting huggingface-hub<1.0,>=0.11.0
  Downloading huggingface_hub-0.13.1-py3-none-any.whl (199 kB)
                          199.2/199.2 kB
23.2 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.27 in
/opt/conda/lib/python3.7/site-packages (from transformers>=4.17.0->ktrain)
(4.64.1)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from langdetect->ktrain) (1.16.0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/opt/conda/lib/python3.7/site-packages (from requests->ktrain) (1.26.14)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.7/site-
packages (from requests->ktrain) (3.4)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/conda/lib/python3.7/site-packages (from requests->ktrain) (2.1.1)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.7/site-packages (from requests->ktrain) (2022.12.7)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/opt/conda/lib/python3.7/site-packages (from scikit-learn->ktrain) (3.1.0)
Requirement already satisfied: scipy>=1.1.0 in /opt/conda/lib/python3.7/site-
packages (from scikit-learn->ktrain) (1.7.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/opt/conda/lib/python3.7/site-packages (from huggingface-
hub<1.0,>=0.11.0->transformers>=4.17.0->ktrain) (4.4.0)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.7/site-
packages (from importlib-metadata->transformers>=4.17.0->ktrain) (3.11.0)
Building wheels for collected packages: ktrain, keras_bert, keras-transformer,
keras-embed-sim, keras-layer-normalization, keras-multi-head, keras-pos-embd,
keras-position-wise-feed-forward, keras-self-attention, jieba, langdetect
  Building wheel for ktrain (setup.py) … done
  Created wheel for ktrain: filename=ktrain-0.33.2-py3-none-any.whl
size=25313824
sha256=5895034cb1cdedd1c2b14e5e40f70783746410331d4203d4c83f98ee2415a4e4
  Stored in directory: /home/jupyter/.cache/pip/wheels/2b/3e/e2/ac8b0795ae6c2271
992501a01cfeaf0530951763c3077b8a7b
  Building wheel for keras_bert (setup.py) … done
  Created wheel for keras_bert: filename=keras_bert-0.89.0-py3-none-
any.whl size=33501
sha256=074d386d631db2598de4745b97371bad0a9527e47768c4c998fd6a3fc867432b
  Stored in directory: /home/jupyter/.cache/pip/wheels/e8/03/69/f1e19e8d13692ff5
b8c928a2b2f418d1dcb6b36632460829bd
  Building wheel for keras-transformer (setup.py) … done
  Created wheel for keras-transformer:
filename=keras_transformer-0.40.0-py3-none-any.whl size=12287
sha256=cdaa2b3ec4d3eb231f8c27e8604b73f51eb6650230c0b4c8ad7018270c29f24d
```

```
  Stored in directory: /home/jupyter/.cache/pip/wheels/37/a3/bf/5f13470e6ff570a9
fecc90d4e24ce34d2ee8b0af43c5333fb0
  Building wheel for keras-embed-sim (setup.py) … done
  Created wheel for keras-embed-sim:
filename=keras_embed_sim-0.10.0-py3-none-any.whl size=3944
sha256=97243c24999e54a7ac2ffc649506dc14b5d90ea708f2fbab1d3e2dba85106530
  Stored in directory: /home/jupyter/.cache/pip/wheels/86/9b/da/cb6fd22132e3675c
de24c5b6f45b94671768fb008cc18cd28b
  Building wheel for keras-layer-normalization (setup.py) … done
  Created wheel for keras-layer-normalization:
filename=keras_layer_normalization-0.16.0-py3-none-any.whl size=4655
sha256=c002537b442dd7cdfa1985f3c564ab43dc9125f6c2d0d44a3bb3b1a68d545cf2
  Stored in directory: /home/jupyter/.cache/pip/wheels/41/f3/10/985c450e02ed9288
fbc5145e90e4726ae95399eaa612a55ee2
  Building wheel for keras-multi-head (setup.py) … done
  Created wheel for keras-multi-head:
filename=keras_multi_head-0.29.0-py3-none-any.whl size=14979
sha256=270b2ce034503127a27c3f5eea8c8ca57730c46d0a2a2e832d4b38c8b0d78417
  Stored in directory: /home/jupyter/.cache/pip/wheels/02/39/c6/cea85ac5607211c0
257754802e9fce5ffc26b4f5fcff351d03
  Building wheel for keras-pos-embd (setup.py) … done
  Created wheel for keras-pos-embd:
filename=keras_pos_embd-0.13.0-py3-none-any.whl size=6946
sha256=6b783bb37b7e7e7cdeee628ba5f5175d49e751685665ea7e4236b80b7c27fe1e
  Stored in directory: /home/jupyter/.cache/pip/wheels/d7/b1/1f/f39f885f243122c1
fb6470bda0827f13c67b58700e71225331
  Building wheel for keras-position-wise-feed-forward (setup.py) … done
  Created wheel for keras-position-wise-feed-forward:
filename=keras_position_wise_feed_forward-0.8.0-py3-none-any.whl size=4968
sha256=f03b1dceb635398f3266a6798b1b0c98ba5cdf5682ccc32d0e96a87048296169
  Stored in directory: /home/jupyter/.cache/pip/wheels/51/d9/ab/db6f4394b1167248
c9e66b932025cd713899fd531f17bb6a92
  Building wheel for keras-self-attention (setup.py) … done
  Created wheel for keras-self-attention:
filename=keras_self_attention-0.51.0-py3-none-any.whl size=18892
sha256=dcb2fdb1b7642e7ef47ef5292406cd647e45bba670777c7fd301a22df1b92c1b
  Stored in directory: /home/jupyter/.cache/pip/wheels/cb/26/00/2d79e29156bddf85
b6c2bccecf43fcb024fb935e3d7a933684
  Building wheel for jieba (setup.py) … done
  Created wheel for jieba: filename=jieba-0.42.1-py3-none-any.whl
size=19314458
sha256=da1dde0c02d7a27abd9d730e1db46fcba1c05e8abb577b522f6664fd54fa943e
  Stored in directory: /home/jupyter/.cache/pip/wheels/db/52/18/8bcb952dbe08a07a
d986c94a1ccca7d5cdd02746bd60d3e846
  Building wheel for langdetect (setup.py) … done
  Created wheel for langdetect: filename=langdetect-1.0.9-py3-none-any.whl
size=993225
sha256=debfd16273f2764cc1164f63bd2fcf21ced3c729c0a99a991c45a134cd0683b6
```

```
  Stored in directory: /home/jupyter/.cache/pip/wheels/73/b2/db/0c9b9eb7a44bf85e
c0b42c06ee617d0a0de66840dc0b3248d1
Successfully built ktrain keras_bert keras-transformer keras-embed-sim keras-
layer-normalization keras-multi-head keras-pos-embd keras-position-wise-feed-
forward keras-self-attention jieba langdetect
Installing collected packages: whoosh, tokenizers, sentencepiece, jieba,
cchardet, syntok, langdetect, keras-self-attention, keras-position-wise-feed-
forward, keras-pos-embd, keras-layer-normalization, keras-embed-sim,
fastprogress, chardet, keras-multi-head, huggingface-hub, transformers, keras-
transformer, keras_bert, ktrain
Successfully installed cchardet-2.1.7 chardet-5.1.0 fastprogress-1.0.3
huggingface-hub-0.13.1 jieba-0.42.1 keras-embed-sim-0.10.0 keras-layer-
normalization-0.16.0 keras-multi-head-0.29.0 keras-pos-embd-0.13.0 keras-
position-wise-feed-forward-0.8.0 keras-self-attention-0.51.0 keras-
transformer-0.40.0 keras_bert-0.89.0 ktrain-0.33.2 langdetect-1.0.9
sentencepiece-0.1.97 syntok-1.4.4 tokenizers-0.13.2 transformers-4.26.1
whoosh-2.7.4
```

[4]:
```
# import ktrain
import ktrain
from ktrain import text
```

[4]:
```
!pip install ktrain --upgrade
```

```
Requirement already satisfied: ktrain in /opt/conda/lib/python3.7/site-packages
(0.33.2)
Requirement already satisfied: packaging in /opt/conda/lib/python3.7/site-
packages (from ktrain) (23.0)
Requirement already satisfied: cchardet in /opt/conda/lib/python3.7/site-
packages (from ktrain) (2.1.7)
Requirement already satisfied: syntok>1.3.3 in /opt/conda/lib/python3.7/site-
packages (from ktrain) (1.4.4)
Requirement already satisfied: scikit-learn in /opt/conda/lib/python3.7/site-
packages (from ktrain) (1.0.2)
Requirement already satisfied: sentencepiece in /opt/conda/lib/python3.7/site-
packages (from ktrain) (0.1.97)
Requirement already satisfied: langdetect in /opt/conda/lib/python3.7/site-
packages (from ktrain) (1.0.9)
Requirement already satisfied: whoosh in /opt/conda/lib/python3.7/site-packages
(from ktrain) (2.7.4)
Requirement already satisfied: keras-bert>=0.86.0 in
/opt/conda/lib/python3.7/site-packages (from ktrain) (0.89.0)
Requirement already satisfied: pandas>=1.0.1 in /opt/conda/lib/python3.7/site-
packages (from ktrain) (1.3.5)
Requirement already satisfied: transformers>=4.17.0 in
/opt/conda/lib/python3.7/site-packages (from ktrain) (4.26.1)
Requirement already satisfied: matplotlib>=3.0.0 in
/opt/conda/lib/python3.7/site-packages (from ktrain) (3.5.3)
```

```
Requirement already satisfied: requests in /opt/conda/lib/python3.7/site-
packages (from ktrain) (2.28.2)
Requirement already satisfied: joblib in /opt/conda/lib/python3.7/site-packages
(from ktrain) (1.2.0)
Requirement already satisfied: chardet in /opt/conda/lib/python3.7/site-packages
(from ktrain) (5.1.0)
Requirement already satisfied: jieba in /opt/conda/lib/python3.7/site-packages
(from ktrain) (0.42.1)
Requirement already satisfied: fastprogress>=0.1.21 in
/opt/conda/lib/python3.7/site-packages (from ktrain) (1.0.3)
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages
(from keras-bert>=0.86.0->ktrain) (1.21.6)
Requirement already satisfied: keras-transformer==0.40.0 in
/opt/conda/lib/python3.7/site-packages (from keras-bert>=0.86.0->ktrain)
(0.40.0)
Requirement already satisfied: keras-position-wise-feed-forward==0.8.0 in
/opt/conda/lib/python3.7/site-packages (from keras-transformer==0.40.0->keras-
bert>=0.86.0->ktrain) (0.8.0)
Requirement already satisfied: keras-pos-embd==0.13.0 in
/opt/conda/lib/python3.7/site-packages (from keras-transformer==0.40.0->keras-
bert>=0.86.0->ktrain) (0.13.0)
Requirement already satisfied: keras-embed-sim==0.10.0 in
/opt/conda/lib/python3.7/site-packages (from keras-transformer==0.40.0->keras-
bert>=0.86.0->ktrain) (0.10.0)
Requirement already satisfied: keras-multi-head==0.29.0 in
/opt/conda/lib/python3.7/site-packages (from keras-transformer==0.40.0->keras-
bert>=0.86.0->ktrain) (0.29.0)
Requirement already satisfied: keras-layer-normalization==0.16.0 in
/opt/conda/lib/python3.7/site-packages (from keras-transformer==0.40.0->keras-
bert>=0.86.0->ktrain) (0.16.0)
Requirement already satisfied: keras-self-attention==0.51.0 in
/opt/conda/lib/python3.7/site-packages (from keras-multi-head==0.29.0->keras-
transformer==0.40.0->keras-bert>=0.86.0->ktrain) (0.51.0)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.7/site-
packages (from matplotlib>=3.0.0->ktrain) (0.11.0)
Requirement already satisfied: pillow>=6.2.0 in /opt/conda/lib/python3.7/site-
packages (from matplotlib>=3.0.0->ktrain) (9.4.0)
Requirement already satisfied: fonttools>=4.22.0 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->ktrain) (4.38.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->ktrain) (1.4.4)
Requirement already satisfied: pyparsing>=2.2.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->ktrain) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=3.0.0->ktrain) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /opt/conda/lib/python3.7/site-
packages (from pandas>=1.0.1->ktrain) (2022.7.1)
Requirement already satisfied: regex>2016 in /opt/conda/lib/python3.7/site-
```

```
packages (from syntok>1.3.3->ktrain) (2022.10.31)
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in
/opt/conda/lib/python3.7/site-packages (from transformers>=4.17.0->ktrain)
(0.13.2)
Requirement already satisfied: huggingface-hub<1.0,>=0.11.0 in
/opt/conda/lib/python3.7/site-packages (from transformers>=4.17.0->ktrain)
(0.13.1)
Requirement already satisfied: pyyaml>=5.1 in /opt/conda/lib/python3.7/site-
packages (from transformers>=4.17.0->ktrain) (6.0)
Requirement already satisfied: filelock in /opt/conda/lib/python3.7/site-
packages (from transformers>=4.17.0->ktrain) (3.9.0)
Requirement already satisfied: importlib-metadata in
/opt/conda/lib/python3.7/site-packages (from transformers>=4.17.0->ktrain)
(6.0.0)
Requirement already satisfied: tqdm>=4.27 in /opt/conda/lib/python3.7/site-
packages (from transformers>=4.17.0->ktrain) (4.64.1)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from langdetect->ktrain) (1.16.0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/opt/conda/lib/python3.7/site-packages (from requests->ktrain) (1.26.14)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.7/site-
packages (from requests->ktrain) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.7/site-packages (from requests->ktrain) (2022.12.7)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/conda/lib/python3.7/site-packages (from requests->ktrain) (2.1.1)
Requirement already satisfied: scipy>=1.1.0 in /opt/conda/lib/python3.7/site-
packages (from scikit-learn->ktrain) (1.7.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/opt/conda/lib/python3.7/site-packages (from scikit-learn->ktrain) (3.1.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/opt/conda/lib/python3.7/site-packages (from huggingface-
hub<1.0,>=0.11.0->transformers>=4.17.0->ktrain) (4.4.0)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.7/site-
packages (from importlib-metadata->transformers>=4.17.0->ktrain) (3.11.0)
```

[5]:
```
ktrain.__version__
```

[5]:
```
'0.33.2'
```

### Check available text classifiers in ktrain

[4]:
```
text.print_text_classifiers()
```

```
fasttext: a fastText-like model [http://arxiv.org/pdf/1607.01759.pdf]
logreg: logistic regression using a trainable Embedding layer
nbsvm: NBSVM model [http://www.aclweb.org/anthology/P12-2018]
bigru: Bidirectional GRU with pretrained fasttext word vectors
```

[https://fasttext.cc/docs/en/crawl-vectors.html]
standard_gru: simple 2-layer GRU with randomly initialized embeddings
bert: Bidirectional Encoder Representations from Transformers (BERT) from keras_bert [https://arxiv.org/abs/1810.04805]
distilbert: distilled, smaller, and faster BERT from Hugging Face transformers [https://arxiv.org/abs/1910.01108]

**Copy files to local FS from GCP bucket**

[5]:
```
!mkdir -p /home/jupyter/data/yelp
```

[6]:
```
!gsutil cp -n 'gs://msca-bdp-data-open/yelp/yelp_train_sentiment.json' '/home/
↪jupyter/data/yelp/'
```

```
Skipping existing item: file:///home/jupyter/data/yelp/yelp_train_sentiment.json
```

### 0.1.1 Load Data

[7]:
```
yelp_path = '/home/jupyter/data/yelp/yelp_train_sentiment.json'
```

[8]:
```
yelp = pd.read_json(yelp_path, orient='records', lines=True)
# yelp = pd.read_json(yelp_path, orient='records', lines=True).head(10000)
yelp.shape
```

[8]:
```
(255717, 3)
```

[9]:
```
yelp.head(5)
```

[9]:

| | text | label | lang |
|---|---|---|---|
| 0 | I love Deagan's. I do. I really do. The atmosp… | 1 | en |
| 1 | I love the classes at this gym. Zumba and. Rad… | 1 | en |
| 2 | The tables and floor were dirty. I was the onl… | 0 | en |
| 3 | I had an oil change at the 15515 N Scottsdale … | 0 | en |
| 4 | The absolute WORST apartment complex I have ev… | 0 | en |

### 0.1.2 Prepare source data

[10]:
```
sentiment = {0: "Negative", 1: "Positive"}
yelp['sentiment'] = yelp['label'].map(sentiment)
```

[11]:
```
df = yelp[['text', 'sentiment']].rename(columns={'text':'data', 'sentiment':
↪'target'})
```

[12]:
```
df.head(5)
```

[12]:

| | data | target |
|---|---|---|
| 0 | I love Deagan's. I do. I really do. The atmosp… | Positive |
| 1 | I love the classes at this gym. Zumba and. Rad… | Positive |

```
2  The tables and floor were dirty. I was the onl…  Negative
3  I had an oil change at the 15515 N Scottsdale …  Negative
4  The absolute WORST apartment complex I have ev…  Negative
```

[13]:
```
df.shape
```

[13]:
```
(255717, 2)
```

[14]:
```
df.groupby(['target']).count()
```

[14]:
```
            data
target
Negative  127995
Positive  127722
```

### 0.2 STEP 1: Load and Preprocess the Dataset

Preprocess the data using the texts_from_array function (since the data resides in an array). If your documents are stored in folders or a CSV file you can use the texts_from_folder or texts_from_csv functions, respectively.

[15]:
```
maxFeatures = 20000 #num of words to consider in vocabulary
maxLen = 200 #each document can be of most <maxLen> words. 0 is used as padding␣
 ↪ID.
nGramRange = 1 #size of multi-word phrases to consider
preprocessMode='standard' #Either 'standard' (normal tokenization) or 'bert'␣
 ↪tokenization and preprocessing for use with BERT text classification model.
sampleSize = 0.3 #Proportion of training to use for validation

(x_train,  y_train), (x_test, y_test), preproc = text.texts_from_df(train_df =␣
 ↪df,
                                                                    text_column␣
 ↪ = 'data',
                                                                   ␣
 ↪label_columns = ['target'],
                                                                      ␣
 ↪val_pct=sampleSize,
                                                                      ␣
 ↪preprocess_mode=preprocessMode, #text must be preprocessed in a specific way␣
 ↪for use with BERT
                                                                      ␣
 ↪maxlen=maxLen,
                                                                      ␣
 ↪max_features=maxFeatures)
```

```
['Negative', 'Positive']
        Negative  Positive
```

```
155978       0.0       1.0
207632       0.0       1.0
47633        1.0       0.0
42203        0.0       1.0
132690       1.0       0.0
['Negative', 'Positive']
        Negative  Positive
72649        1.0       0.0
114977       0.0       1.0
248772       1.0       0.0
1180         1.0       0.0
144184       1.0       0.0
language: en
Word Counts: 109584
Nrows: 179001
179001 train sequences
train sequence lengths:
        mean : 113
        95percentile : 327
        99percentile : 578
x_train shape: (179001,200)
y_train shape: (179001, 2)
Is Multi-Label? False
76716 test sequences
test sequence lengths:
        mean : 113
        95percentile : 325
        99percentile : 587
x_test shape: (76716,200)
y_test shape: (76716, 2)
```

### 0.3 STEP 2: Load a pretrained fastText model and wrap it in a `ktrain.Learner` object

This step can be condensed into a single line of code, but we execute it as two lines for clarity. (You can ignore the deprecation warnings arising from Keras 2.2.4 with TensorFlow 1.14.0.)

[16]:
```
model = text.text_classifier('bigru', (x_train, y_train), preproc=preproc)
```

```
Is Multi-Label? False
compiling word ID features…
maxlen is 200
word vectors will be loaded from:
https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.en.300.vec.gz
processing pretrained word vectors…
loading pretrained word vectors…this may take a few moments…
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>

2023-03-10 03:58:22.643585: I tensorflow/core/platform/cpu_feature_guard.cc:193]
This TensorFlow binary is optimized with oneAPI Deep Neural Network Library
(oneDNN) to use the following CPU instructions in performance-critical
operations:  AVX2 FMA
To enable them in other operations, rebuild TensorFlow with the appropriate
compiler flags.

done.
```

[17]:
```
batchSize = 64 ### Check best size

learner = ktrain.get_learner(model,
                             train_data=(x_train, y_train),
                             val_data=(x_test, y_test),
                             batch_size=batchSize)
```

### 0.4 STEP 3: Train the Model

We train using one of the three learning rates recommended in the BERT paper: *5e-5*, *3e-5*, or *2e-5*. Alternatively, the ktrain Learning Rate Finder can be used to find a good learning rate by invoking `learner.lr_find()` and `learner.lr_plot()`, prior to training. The `learner.fit_onecycle` method employs a 1cycle learning rate policy.

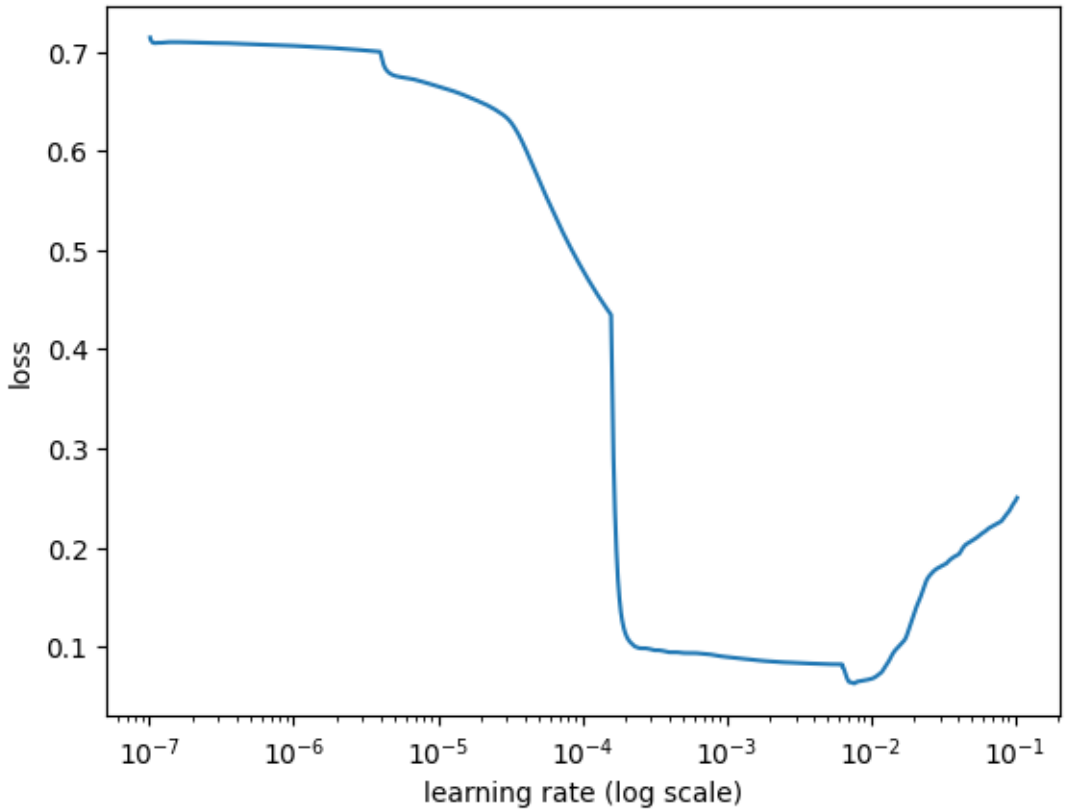
[19]:
```
# briefly simulate training to find good learning rate

learner.lr_find(max_epochs=5)
```

```
simulating training for different learning rates… this may take a few
moments…
Epoch 1/5
2797/2797 [==============================] - 463s 163ms/step - loss: 0.6996 -
accuracy: 0.5097
Epoch 2/5
2797/2797 [==============================] - 463s 165ms/step - loss: 0.4289 -
accuracy: 0.7923
Epoch 3/5
2797/2797 [==============================] - 474s 169ms/step - loss: 0.0822 -
accuracy: 0.9704
Epoch 4/5
2797/2797 [==============================] - 364s 130ms/step - loss: 0.2639 -
accuracy: 0.9203


done.
Please invoke the Learner.lr_plot() method to visually inspect the loss plot to
help identify the maximal learning rate associated with falling loss.
```

[20]:
```
learner.lr_plot()
```




[18]:
```
learningRate = 0.001
numEpoch = 3

# learner.fit_onecycle(learningRate, numEpoch)
learner.autofit(learningRate, numEpoch)
```

```
begin training using triangular learning rate policy with max lr of 0.001…
Epoch 1/3
2797/2797 [==============================] - 548s 194ms/step - loss: 0.1222 -
accuracy: 0.9486 - val_loss: 0.0516 - val_accuracy: 0.9820
Epoch 2/3
2797/2797 [==============================] - 546s 195ms/step - loss: 0.0420 -
accuracy: 0.9857 - val_loss: 0.0479 - val_accuracy: 0.9831
Epoch 3/3
2797/2797 [==============================] - 549s 196ms/step - loss: 0.0269 -
accuracy: 0.9915 - val_loss: 0.0492 - val_accuracy: 0.9831
```

[18]:
```
<keras.callbacks.History at 0x7f2fb1bc1850>
```

We can use the `learner.validate` method to test our model against the validation set.

[19]:
```
learner.validate(val_data=(x_test, y_test))
```

```
2398/2398 [==============================] - 87s 36ms/step
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     38094
           1       0.98      0.98      0.98     38622

    accuracy                           0.98     76716
   macro avg       0.98      0.98      0.98     76716
weighted avg       0.98      0.98      0.98     76716
```

[19]:
```
array([[37507,   587],
       [  707, 37915]])
```

[20]:
```
learner.validate(val_data=(x_test, y_test))
```

```
2398/2398 [==============================] - 86s 36ms/step
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     38094
           1       0.98      0.98      0.98     38622

    accuracy                           0.98     76716
   macro avg       0.98      0.98      0.98     76716
weighted avg       0.98      0.98      0.98     76716
```

[20]:
```
array([[37507,   587],
       [  707, 37915]])
```

### 0.5 STEP 4: Making predictions

We can call the `learner.get_predictor` method to obtain a Predictor object capable of making predictions on new raw data.

[21]:
```
predictor = ktrain.get_predictor(learner.model, preproc)
```

[22]:
```
predictor.get_classes()
```

[22]:
```
['Negative', 'Positive']
```

[23]:
```
df=pd.read_parquet("news_df_TM.parquet")
df.shape
```

[23]:
```
(199838, 8)
```

[24]:
```
df.head(2)
```

[24]:
```
   index                                                url        date  \
0      0  http://auckland.scoop.co.nz/2020/01/aut-boosts…  2020-01-28
1      1  http://en.people.cn/n3/2021/0318/c90000-983012…  2021-03-18

  language                                              title  \
0       en  auckland.scoop.co.nz » AUT boosts AI expertise…
1       en  Artificial intelligence improves parking effic…

                                                 text  \
0  \n\nauckland.scoop.co.nz » AUT boosts AI exper…
1  \n\nArtificial intelligence improves parking e…

                                      text_cleaned  topic
0  aucklandscoopconz aut boost ai expertise new a…      2
1  artificial intelligence improves parking effic…      3
```

[26]:
```
print(predictor.predict(df.text_cleaned.iloc[0]))
```

```
1/1 [==============================] - 0s 106ms/step
Positive
```

[28]:
```
NumRecs = len(df)

#target = df.text_cleaned.iloc[0:NumRecs]
predicted = predictor.predict(df.text_cleaned.iloc[0:NumRecs].tolist())
#data = df.text_cleaned.iloc[0:NumRecs]

#results = pd.DataFrame(list(zip(target, predicted, data)),
               #columns =['target', 'predicted', 'data'])
len(predicted)
```

```
6245/6245 [==============================] - 219s 35ms/step
```

[28]:
```
199838
```

[29]:
```
df["Sentiment"]=predicted
```

[30]:
```
df.head(2)
```

[30]:
```
   index                                                url        date  \
0      0  http://auckland.scoop.co.nz/2020/01/aut-boosts…  2020-01-28
1      1  http://en.people.cn/n3/2021/0318/c90000-983012…  2021-03-18

  language                                              title  \
0       en  auckland.scoop.co.nz » AUT boosts AI expertise…
1       en  Artificial intelligence improves parking effic…

                                                  text  \
0  \n\nauckland.scoop.co.nz » AUT boosts AI exper…
1  \n\nArtificial intelligence improves parking e…

                                       text_cleaned  topic Sentiment
0  aucklandscoopconz aut boost ai expertise new a…      2  Positive
1  artificial intelligence improves parking effic…      3  Negative
```

[38]:
```
df.Sentiment.value_counts()
```

[38]:
```
Negative    161709
Positive     38129
Name: Sentiment, dtype: int64
```

[33]:
```
df['date'] = pd.to_datetime(df['date'])
```

[34]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 199838 entries, 0 to 199837
Data columns (total 9 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   index         199838 non-null  int64
 1   url           199838 non-null  object
 2   date          199838 non-null  datetime64[ns]
 3   language      199838 non-null  object
 4   title         199838 non-null  object
 5   text          199838 non-null  object
 6   text_cleaned  199838 non-null  object
 7   topic         199838 non-null  int64
 8   Sentiment     199838 non-null  object
dtypes: datetime64[ns](1), int64(2), object(6)
memory usage: 13.7+ MB
```

[53]:
```
df["topic"]=df["topic"]+1
```

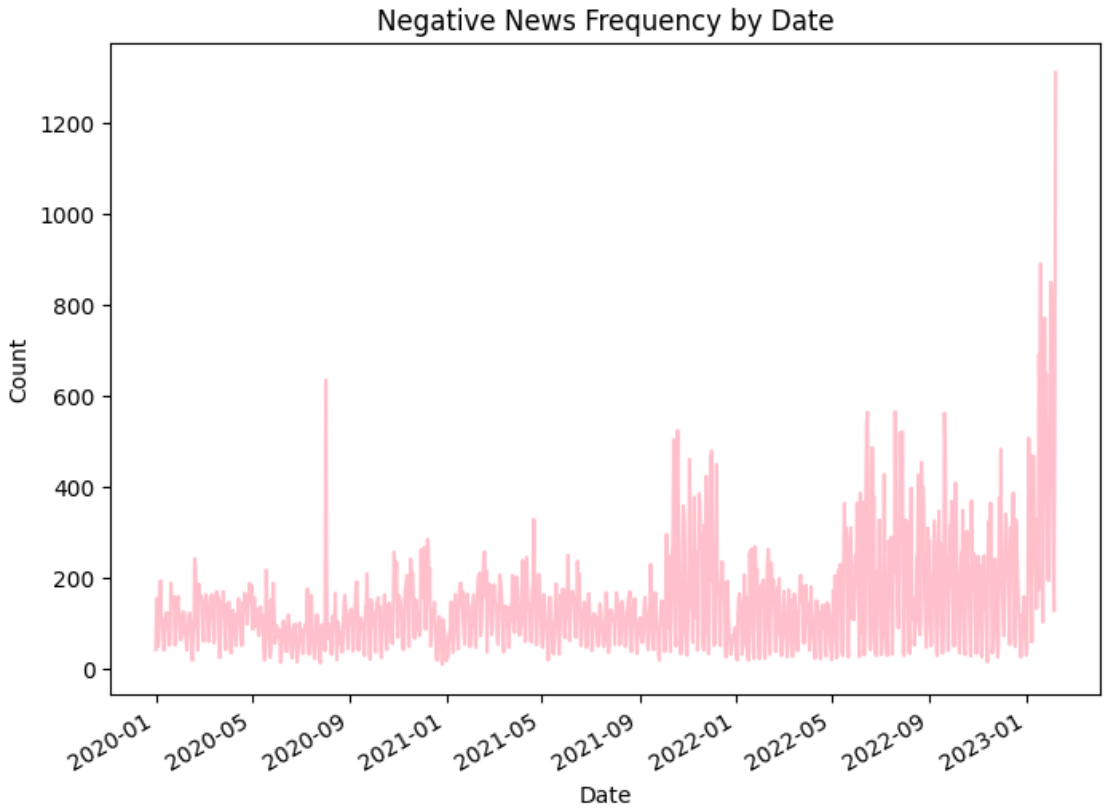
[54]:
```
# group by date and count the occurrences of 'column_to_count'
NegSent_df=df[df["Sentiment"]=='Negative']
```

```
NegSentiment_by_date = NegSent_df.groupby('date')['Sentiment'].count()
NegSentiment_by_date
```

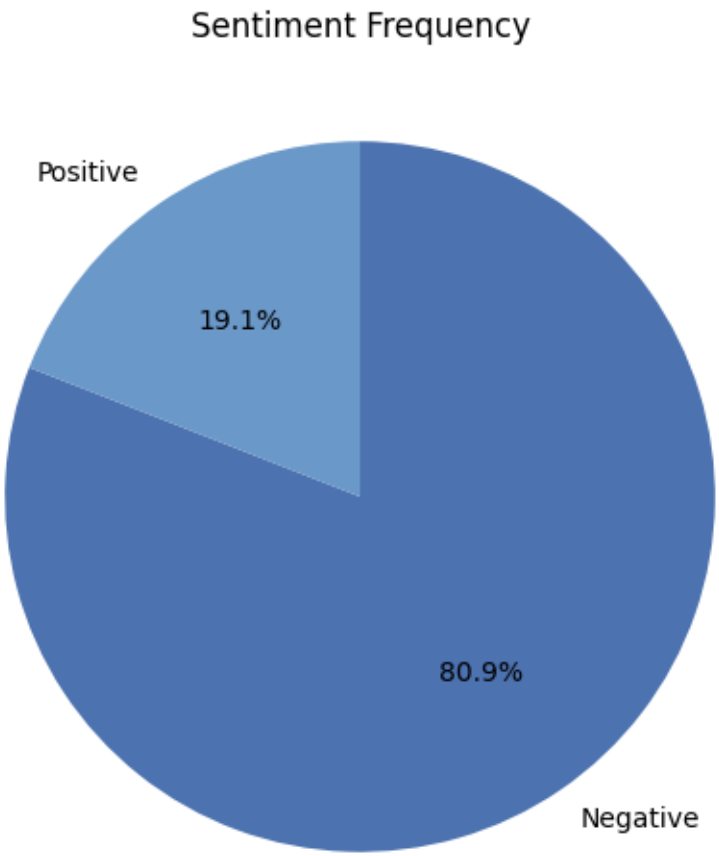
[54]:
```
date
2020-01-01      41
2020-01-02     138
2020-01-03     155
2020-01-04      48
2020-01-05      73
               …
2023-02-03     420
2023-02-04     213
2023-02-05     127
2023-02-06     636
2023-02-07    1312
Name: Sentiment, Length: 1133, dtype: int64
```

[49]:
```
import matplotlib.pyplot as plt

# plot a line chart of the count by date
NegSentiment_by_date.plot(kind='line', figsize=(8,6),color='pink')
plt.title('Negative News Frequency by Date')
plt.xlabel('Date')
plt.ylabel('Count')
plt.show()
```




[51]:
```
# create a pie chart of the value counts of 'column_to_count'
color_palette = ['#4c72b0', '#6a98c9', '#9cd1fc']
df['Sentiment'].value_counts().plot(kind='pie', figsize=(8,6), autopct='%1.
  ↪1f%%', startangle=90, counterclock=False,colors=color_palette)
plt.title('Sentiment Frequency')
plt.ylabel('')
plt.show()
```

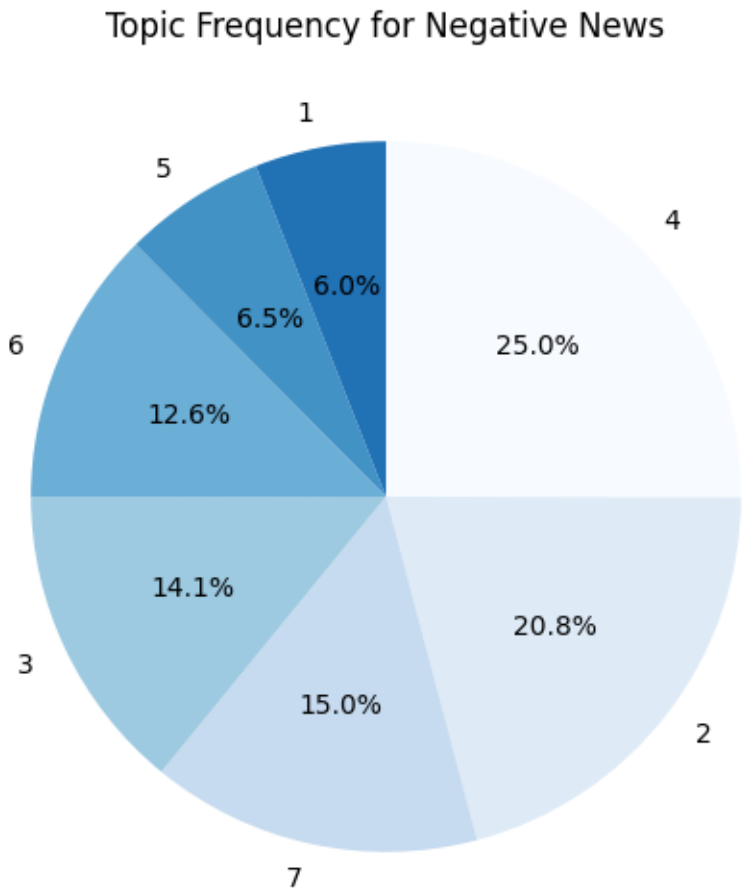



[ ]:
```
temp=df[df["Sentiment"]=='Negative']
```

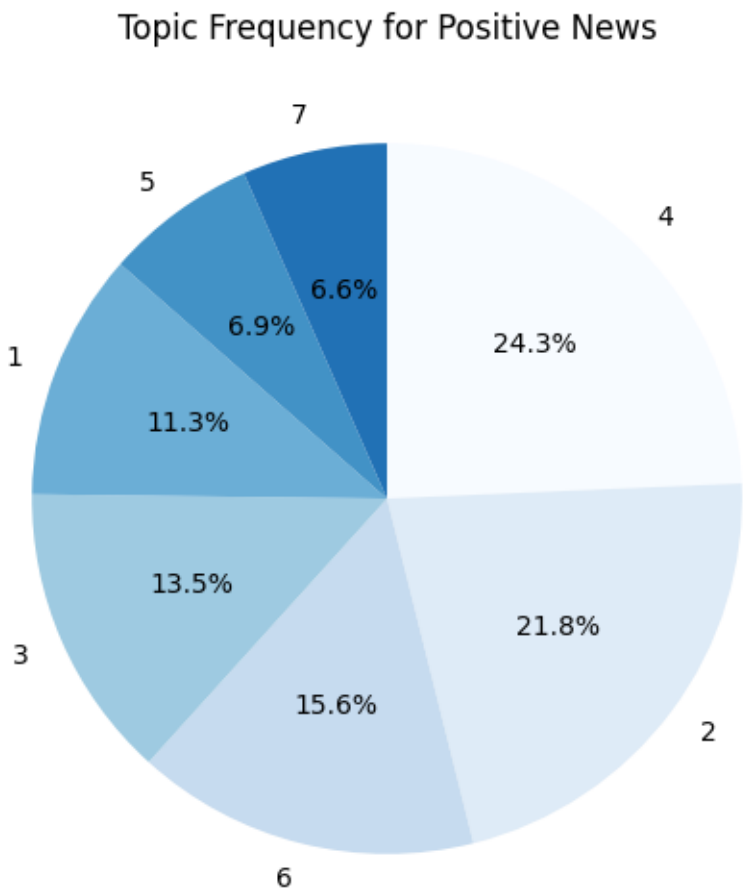
[66]:
```
from matplotlib import colors

# create a blue shades color palette with 7 colors
color_palette = colors.ListedColormap(['#f7fbff', '#deebf7', '#c6dbef',␣
 ↪'#9ecae1', '#6baed6', '#4292c6', '#2171b5'])

# create a pie chart of the value counts of 'column_to_count' using the blue␣
 ↪shades color palette
NegSent_df['topic'].value_counts().plot(kind='pie', figsize=(8,6), autopct='%1.
 ↪1f%%', startangle=90, counterclock=False, colors=color_palette.colors)
plt.title('Topic Frequency for Negative News')
plt.ylabel('')
plt.show()
```



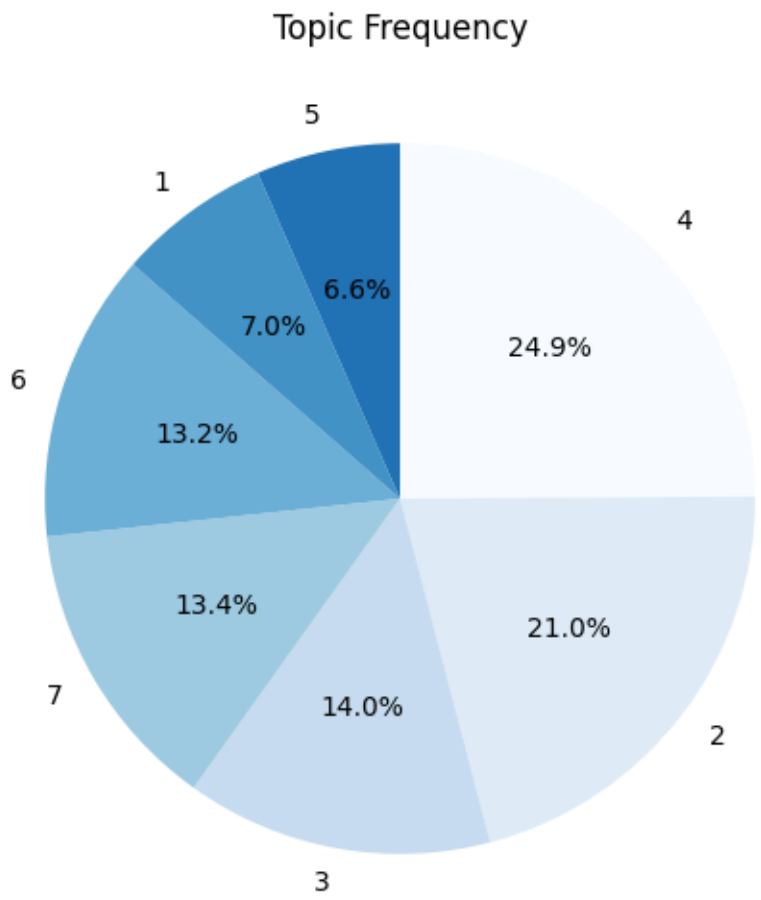

[68]:
```
PosSent_df=df[df["Sentiment"]=='Positive']
# create a pie chart of the value counts of 'column_to_count' using the blue␣
 ↪shades color palette
PosSent_df['topic'].value_counts().plot(kind='pie', figsize=(8,6), autopct='%1.
 ↪1f%%', startangle=90, counterclock=False, colors=color_palette.colors)
plt.title('Topic Frequency for Positive News')
plt.ylabel('')
plt.show()
```




[5]:
```
import matplotlib.pyplot as plt
from matplotlib import colors

# create a blue shades color palette with 7 colors
color_palette = colors.ListedColormap(['#f7fbff', '#deebf7', '#c6dbef',␣
 ↪'#9ecae1', '#6baed6', '#4292c6', '#2171b5'])

# create a pie chart of the value counts of 'column_to_count' using the blue␣
 ↪shades color palette
df['topic'].value_counts().plot(kind='pie', figsize=(8,6), autopct='%1.1f%%',␣
 ↪startangle=90, counterclock=False, colors=color_palette.colors)
plt.title('Topic Frequency')
plt.ylabel('')
plt.show()
```



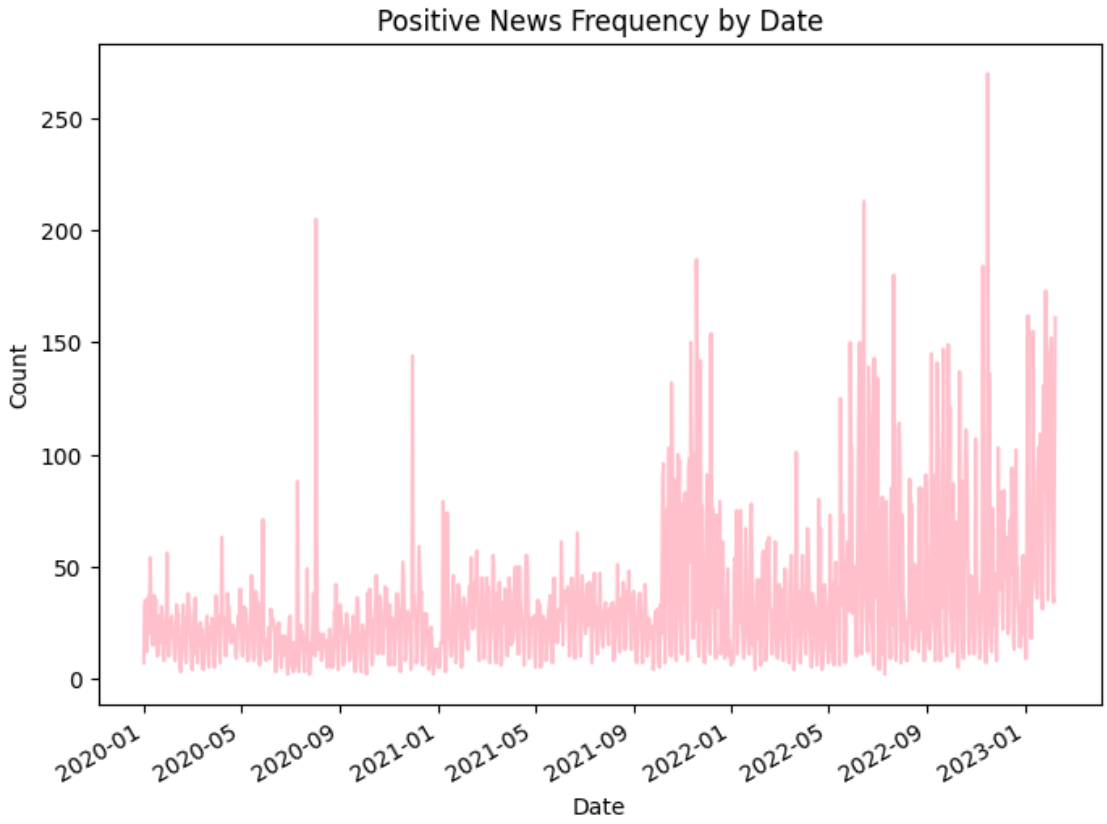

[2]:
```
# group by date and count the occurrences of 'column_to_count'
PosSent_df=df[df["Sentiment"]=='Positive']
PosSentiment_by_date = PosSent_df.groupby('date')['Sentiment'].count()
PosSentiment_by_date
```

[2]:
```
date
2020-01-01      7
2020-01-02     32
2020-01-03     35
2020-01-04     16
2020-01-05     12
              ...
2023-02-03     85
2023-02-04     38
2023-02-05     34
2023-02-06     98
2023-02-07    161
```

```
Name: Sentiment, Length: 1132, dtype: int64
```

[3]:
```
import matplotlib.pyplot as plt

# plot a line chart of the count by date
PosSentiment_by_date.plot(kind='line', figsize=(8,6),color='pink')
plt.title('Positive News Frequency by Date')
plt.xlabel('Date')
plt.ylabel('Count')
plt.show()
```




[32]:
```
df.to_parquet("news_v3.parquet")
```

[1]:
```
import pandas as pd
df=pd.read_parquet("news_v3.parquet")
df.shape
```

[1]:
```
(199838, 9)
```

[4]:
```
df.head(2)
```

[4]:

```
   index                                                url       date  \
0      0  http://auckland.scoop.co.nz/2020/01/aut-boosts… 2020-01-28
1      1  http://en.people.cn/n3/2021/0318/c90000-983012… 2021-03-18

  language                                              title  \
0       en  auckland.scoop.co.nz » AUT boosts AI expertise…
1       en  Artificial intelligence improves parking effic…

                                                  text  \
0  \n\nauckland.scoop.co.nz » AUT boosts AI exper…
1  \n\nArtificial intelligence improves parking e…

                                       text_cleaned  topic Sentiment
0  aucklandscoopconz aut boost ai expertise new a…      3  Positive
1  artificial intelligence improves parking effic…      4  Negative
```

[ ]: