text-analytics-group5

March 29, 2025

[1]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[2]:
```
!pip install pandas numpy nltk matplotlib wordcloud seaborn openpyxl
!pip install datasets
!pip install evaluate
!pip install transformers
!pip install --upgrade pip
!pip install datasets scikit-learn transformers evaluate huggingface_hub
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(2.2.2)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages
(2.0.2)
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages
(3.9.1)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.11/dist-
packages (3.10.0)
Requirement already satisfied: wordcloud in /usr/local/lib/python3.11/dist-
packages (1.9.4)
Requirement already satisfied: seaborn in /usr/local/lib/python3.11/dist-
packages (0.13.2)
Requirement already satisfied: openpyxl in /usr/local/lib/python3.11/dist-
packages (3.1.5)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas) (2025.1)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages
(from nltk) (8.1.8)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages
(from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in
/usr/local/lib/python3.11/dist-packages (from nltk) (2024.11.6)
```

```
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages
(from nltk) (4.67.1)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-
packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-
packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (3.2.1)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.11/dist-
packages (from openpyxl) (2.0.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
Collecting datasets
  Downloading datasets-3.5.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.18.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from datasets) (2.0.2)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(from datasets) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in
/usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.11/dist-
packages (from datasets) (4.67.1)
Collecting xxhash (from datasets)
  Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(12 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2 kB)
Collecting fsspec<=2024.12.0,>=2023.1.0 (from
fsspec[http]<=2024.12.0,>=2023.1.0->datasets)
  Downloading fsspec-2024.12.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.11.14)
Requirement already satisfied: huggingface-hub>=0.24.0 in
```

```
/usr/local/lib/python3.11/dist-packages (from datasets) (0.29.3)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from datasets) (6.0.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.2.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.18.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.24.0->datasets)
(4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests>=2.32.2->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2025.1.31)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->datasets) (1.17.0)
Downloading datasets-3.5.0-py3-none-any.whl (491 kB)
                        491.2/491.2 kB
14.5 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                        116.3/116.3 kB
8.5 MB/s eta 0:00:00
Downloading fsspec-2024.12.0-py3-none-any.whl (183 kB)
```

```
                        183.9/183.9 kB
14.0 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py311-none-any.whl (143 kB)
                        143.5/143.5 kB
9.0 MB/s eta 0:00:00
Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
                        194.8/194.8 kB
13.8 MB/s eta 0:00:00
Installing collected packages: xxhash, fsspec, dill, multiprocess,
datasets
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2025.3.0
    Uninstalling fsspec-2025.3.0:
      Successfully uninstalled fsspec-2025.3.0
```

```
ERROR: pip's dependency resolver does not currently take into account all
the packages that are installed. This behaviour is the source of the following
dependency conflicts.
gcsfs 2025.3.0 requires fsspec==2025.3.0, but you have fsspec 2024.12.0 which is
incompatible.
torch 2.6.0+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cublas-cu12
12.5.3.2 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-cupti-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cuda-cupti-cu12
12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-nvrtc-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cuda-nvrtc-cu12
12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-runtime-cu12==12.4.127; platform_system
== "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-runtime-
cu12 12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cudnn-cu12==9.1.0.70; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cudnn-cu12
9.3.0.75 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cufft-cu12==11.2.1.3; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cufft-cu12
11.2.3.61 which is incompatible.
torch 2.6.0+cu124 requires nvidia-curand-cu12==10.3.5.147; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-curand-cu12
10.3.6.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cusolver-cu12==11.6.1.9; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cusolver-cu12
11.6.3.83 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cusparse-cu12==12.3.1.170; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cusparse-cu12
12.5.1.3 which is incompatible.
torch 2.6.0+cu124 requires nvidia-nvjitlink-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-nvjitlink-cu12
12.5.82 which is incompatible.
Successfully installed datasets-3.5.0 dill-0.3.8 fsspec-2024.12.0
```

```
multiprocess-0.70.16 xxhash-3.5.0
Collecting evaluate
  Downloading evaluate-0.4.3-py3-none-any.whl.metadata (9.2 kB)
Requirement already satisfied: datasets>=2.0.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (3.5.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from evaluate) (2.0.2)
Requirement already satisfied: dill in /usr/local/lib/python3.11/dist-packages
(from evaluate) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(from evaluate) (2.2.2)
Requirement already satisfied: requests>=2.19.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (2.32.3)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.11/dist-
packages (from evaluate) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages
(from evaluate) (3.5.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.11/dist-
packages (from evaluate) (0.70.16)
Requirement already satisfied: fsspec>=2021.05.0 in
/usr/local/lib/python3.11/dist-packages (from fsspec[http]>=2021.05.0->evaluate)
(2024.12.0)
Requirement already satisfied: huggingface-hub>=0.7.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (0.29.3)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from evaluate) (24.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (3.18.0)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.0.0->evaluate)
(18.1.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (3.11.14)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (6.0.2)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.7.0->evaluate)
(4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests>=2.19.0->evaluate) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
```

```
(2025.1.31)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->evaluate) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->evaluate) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->evaluate) (2025.1)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->datasets>=2.0.0->evaluate) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (6.2.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.18.3)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->evaluate) (1.17.0)
Downloading evaluate-0.4.3-py3-none-any.whl (84 kB)
                        84.0/84.0 kB
3.1 MB/s eta 0:00:00
Installing collected packages: evaluate
Successfully installed evaluate-0.4.3
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-
packages (4.50.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.29.3)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
```

```
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-
packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.3 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-
packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.11/dist-packages (from huggingface-
hub<1.0,>=0.26.0->transformers) (2024.12.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-
hub<1.0,>=0.26.0->transformers) (4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(2025.1.31)
Requirement already satisfied: pip in /usr/local/lib/python3.11/dist-packages
(24.1.2)
Collecting pip
  Downloading pip-25.0.1-py3-none-any.whl.metadata (3.7 kB)
Downloading pip-25.0.1-py3-none-any.whl (1.8 MB)
                         1.8/1.8 MB
29.8 MB/s eta 0:00:00
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 24.1.2
    Uninstalling pip-24.1.2:
      Successfully uninstalled pip-24.1.2
Successfully installed pip-25.0.1
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-
packages (3.5.0)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-
packages (1.6.1)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-
packages (4.50.0)
Requirement already satisfied: evaluate in /usr/local/lib/python3.11/dist-
packages (0.4.3)
Requirement already satisfied: huggingface_hub in
/usr/local/lib/python3.11/dist-packages (0.29.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.18.0)
```

```
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from datasets) (2.0.2)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(from datasets) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in
/usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.11/dist-
packages (from datasets) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages
(from datasets) (3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in
/usr/local/lib/python3.11/dist-packages (from datasets) (0.70.16)
Requirement already satisfied: fsspec<=2024.12.0,>=2023.1.0 in
/usr/local/lib/python3.11/dist-packages (from
fsspec[http]<=2024.12.0,>=2023.1.0->datasets) (2024.12.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.11.14)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from datasets) (6.0.2)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn) (1.14.1)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn) (3.6.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.3 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface_hub) (4.12.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
```

```
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.2.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.18.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests>=2.32.2->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2025.1.31)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->datasets) (1.17.0)
```

[8]:
```python
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
from wordcloud import STOPWORDS
from collections import Counter
import re
import nltk
nltk.download('all')  # Download all available resources (fixes any missing␣
 ↪ones)
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer

# Download NLTK resources
nltk.download("stopwords")
nltk.download("punkt")
nltk.download("wordnet")
```

```
[nltk_data] Downloading collection 'all'
[nltk_data]    |
```

```
[nltk_data]     | Downloading package abc to /root/nltk_data…
[nltk_data]     |   Package abc is already up-to-date!
[nltk_data]     | Downloading package alpino to /root/nltk_data…
[nltk_data]     |   Package alpino is already up-to-date!
[nltk_data]     | Downloading package averaged_perceptron_tagger to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package averaged_perceptron_tagger is already up-
[nltk_data]     |       to-date!
[nltk_data]     | Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package averaged_perceptron_tagger_eng is already
[nltk_data]     |       up-to-date!
[nltk_data]     | Downloading package averaged_perceptron_tagger_ru to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package averaged_perceptron_tagger_ru is already
[nltk_data]     |       up-to-date!
[nltk_data]     | Downloading package averaged_perceptron_tagger_rus to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package averaged_perceptron_tagger_rus is already
[nltk_data]     |       up-to-date!
[nltk_data]     | Downloading package basque_grammars to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package basque_grammars is already up-to-date!
[nltk_data]     | Downloading package bcp47 to /root/nltk_data…
[nltk_data]     |   Package bcp47 is already up-to-date!
[nltk_data]     | Downloading package biocreative_ppi to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package biocreative_ppi is already up-to-date!
[nltk_data]     | Downloading package bllip_wsj_no_aux to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package bllip_wsj_no_aux is already up-to-date!
[nltk_data]     | Downloading package book_grammars to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package book_grammars is already up-to-date!
[nltk_data]     | Downloading package brown to /root/nltk_data…
[nltk_data]     |   Package brown is already up-to-date!
[nltk_data]     | Downloading package brown_tei to /root/nltk_data…
[nltk_data]     |   Package brown_tei is already up-to-date!
[nltk_data]     | Downloading package cess_cat to /root/nltk_data…
[nltk_data]     |   Package cess_cat is already up-to-date!
[nltk_data]     | Downloading package cess_esp to /root/nltk_data…
[nltk_data]     |   Package cess_esp is already up-to-date!
[nltk_data]     | Downloading package chat80 to /root/nltk_data…
[nltk_data]     |   Package chat80 is already up-to-date!
[nltk_data]     | Downloading package city_database to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package city_database is already up-to-date!
[nltk_data]     | Downloading package cmudict to /root/nltk_data…
```

```
[nltk_data]    |   Package cmudict is already up-to-date!
[nltk_data]    | Downloading package comparative_sentences to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package comparative_sentences is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package comtrans to /root/nltk_data…
[nltk_data]    |   Package comtrans is already up-to-date!
[nltk_data]    | Downloading package conll2000 to /root/nltk_data…
[nltk_data]    |   Package conll2000 is already up-to-date!
[nltk_data]    | Downloading package conll2002 to /root/nltk_data…
[nltk_data]    |   Package conll2002 is already up-to-date!
[nltk_data]    | Downloading package conll2007 to /root/nltk_data…
[nltk_data]    |   Package conll2007 is already up-to-date!
[nltk_data]    | Downloading package crubadan to /root/nltk_data…
[nltk_data]    |   Package crubadan is already up-to-date!
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package dependency_treebank is already up-to-date!
[nltk_data]    | Downloading package dolch to /root/nltk_data…
[nltk_data]    |   Package dolch is already up-to-date!
[nltk_data]    | Downloading package english_wordnet to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package english_wordnet is already up-to-date!
[nltk_data]    | Downloading package europarl_raw to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package europarl_raw is already up-to-date!
[nltk_data]    | Downloading package extended_omw to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package extended_omw is already up-to-date!
[nltk_data]    | Downloading package floresta to /root/nltk_data…
[nltk_data]    |   Package floresta is already up-to-date!
[nltk_data]    | Downloading package framenet_v15 to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package framenet_v15 is already up-to-date!
[nltk_data]    | Downloading package framenet_v17 to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package framenet_v17 is already up-to-date!
[nltk_data]    | Downloading package gazetteers to /root/nltk_data…
[nltk_data]    |   Package gazetteers is already up-to-date!
[nltk_data]    | Downloading package genesis to /root/nltk_data…
[nltk_data]    |   Package genesis is already up-to-date!
[nltk_data]    | Downloading package gutenberg to /root/nltk_data…
[nltk_data]    |   Package gutenberg is already up-to-date!
[nltk_data]    | Downloading package ieer to /root/nltk_data…
[nltk_data]    |   Package ieer is already up-to-date!
[nltk_data]    | Downloading package inaugural to /root/nltk_data…
[nltk_data]    |   Package inaugural is already up-to-date!
[nltk_data]    | Downloading package indian to /root/nltk_data…
```

```
[nltk_data]    |   Package indian is already up-to-date!
[nltk_data]    | Downloading package jeita to /root/nltk_data…
[nltk_data]    |   Package jeita is already up-to-date!
[nltk_data]    | Downloading package kimmo to /root/nltk_data…
[nltk_data]    |   Package kimmo is already up-to-date!
[nltk_data]    | Downloading package knbc to /root/nltk_data…
[nltk_data]    |   Package knbc is already up-to-date!
[nltk_data]    | Downloading package large_grammars to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package large_grammars is already up-to-date!
[nltk_data]    | Downloading package lin_thesaurus to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package lin_thesaurus is already up-to-date!
[nltk_data]    | Downloading package mac_morpho to /root/nltk_data…
[nltk_data]    |   Package mac_morpho is already up-to-date!
[nltk_data]    | Downloading package machado to /root/nltk_data…
[nltk_data]    |   Package machado is already up-to-date!
[nltk_data]    | Downloading package masc_tagged to /root/nltk_data…
[nltk_data]    |   Package masc_tagged is already up-to-date!
[nltk_data]    | Downloading package maxent_ne_chunker to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package maxent_ne_chunker is already up-to-date!
[nltk_data]    | Downloading package maxent_ne_chunker_tab to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package maxent_ne_chunker_tab is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package maxent_treebank_pos_tagger to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package maxent_treebank_pos_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package maxent_treebank_pos_tagger_tab to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package maxent_treebank_pos_tagger_tab is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package moses_sample to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package moses_sample is already up-to-date!
[nltk_data]    | Downloading package movie_reviews to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package movie_reviews is already up-to-date!
[nltk_data]    | Downloading package mte_teip5 to /root/nltk_data…
[nltk_data]    |   Package mte_teip5 is already up-to-date!
[nltk_data]    | Downloading package mwa_ppdb to /root/nltk_data…
[nltk_data]    |   Package mwa_ppdb is already up-to-date!
[nltk_data]    | Downloading package names to /root/nltk_data…
[nltk_data]    |   Package names is already up-to-date!
[nltk_data]    | Downloading package nombank.1.0 to /root/nltk_data…
[nltk_data]    |   Package nombank.1.0 is already up-to-date!
```

```
[nltk_data]    | Downloading package nonbreaking_prefixes to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package nonbreaking_prefixes is already up-to-date!
[nltk_data]    | Downloading package nps_chat to /root/nltk_data…
[nltk_data]    |   Package nps_chat is already up-to-date!
[nltk_data]    | Downloading package omw to /root/nltk_data…
[nltk_data]    |   Package omw is already up-to-date!
[nltk_data]    | Downloading package omw-1.4 to /root/nltk_data…
[nltk_data]    |   Package omw-1.4 is already up-to-date!
[nltk_data]    | Downloading package opinion_lexicon to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package opinion_lexicon is already up-to-date!
[nltk_data]    | Downloading package panlex_swadesh to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package panlex_swadesh is already up-to-date!
[nltk_data]    | Downloading package paradigms to /root/nltk_data…
[nltk_data]    |   Package paradigms is already up-to-date!
[nltk_data]    | Downloading package pe08 to /root/nltk_data…
[nltk_data]    |   Package pe08 is already up-to-date!
[nltk_data]    | Downloading package perluniprops to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package perluniprops is already up-to-date!
[nltk_data]    | Downloading package pil to /root/nltk_data…
[nltk_data]    |   Package pil is already up-to-date!
[nltk_data]    | Downloading package pl196x to /root/nltk_data…
[nltk_data]    |   Package pl196x is already up-to-date!
[nltk_data]    | Downloading package porter_test to /root/nltk_data…
[nltk_data]    |   Package porter_test is already up-to-date!
[nltk_data]    | Downloading package ppattach to /root/nltk_data…
[nltk_data]    |   Package ppattach is already up-to-date!
[nltk_data]    | Downloading package problem_reports to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package problem_reports is already up-to-date!
[nltk_data]    | Downloading package product_reviews_1 to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package product_reviews_1 is already up-to-date!
[nltk_data]    | Downloading package product_reviews_2 to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package product_reviews_2 is already up-to-date!
[nltk_data]    | Downloading package propbank to /root/nltk_data…
[nltk_data]    |   Package propbank is already up-to-date!
[nltk_data]    | Downloading package pros_cons to /root/nltk_data…
[nltk_data]    |   Package pros_cons is already up-to-date!
[nltk_data]    | Downloading package ptb to /root/nltk_data…
[nltk_data]    |   Package ptb is already up-to-date!
[nltk_data]    | Downloading package punkt to /root/nltk_data…
[nltk_data]    |   Package punkt is already up-to-date!
[nltk_data]    | Downloading package punkt_tab to /root/nltk_data…
```

```
[nltk_data]    |   Package punkt_tab is already up-to-date!
[nltk_data]    | Downloading package qc to /root/nltk_data…
[nltk_data]    |   Package qc is already up-to-date!
[nltk_data]    | Downloading package reuters to /root/nltk_data…
[nltk_data]    |   Package reuters is already up-to-date!
[nltk_data]    | Downloading package rslp to /root/nltk_data…
[nltk_data]    |   Package rslp is already up-to-date!
[nltk_data]    | Downloading package rte to /root/nltk_data…
[nltk_data]    |   Package rte is already up-to-date!
[nltk_data]    | Downloading package sample_grammars to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package sample_grammars is already up-to-date!
[nltk_data]    | Downloading package semcor to /root/nltk_data…
[nltk_data]    |   Package semcor is already up-to-date!
[nltk_data]    | Downloading package senseval to /root/nltk_data…
[nltk_data]    |   Package senseval is already up-to-date!
[nltk_data]    | Downloading package sentence_polarity to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package sentence_polarity is already up-to-date!
[nltk_data]    | Downloading package sentiwordnet to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package sentiwordnet is already up-to-date!
[nltk_data]    | Downloading package shakespeare to /root/nltk_data…
[nltk_data]    |   Package shakespeare is already up-to-date!
[nltk_data]    | Downloading package sinica_treebank to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package sinica_treebank is already up-to-date!
[nltk_data]    | Downloading package smultron to /root/nltk_data…
[nltk_data]    |   Package smultron is already up-to-date!
[nltk_data]    | Downloading package snowball_data to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package snowball_data is already up-to-date!
[nltk_data]    | Downloading package spanish_grammars to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package spanish_grammars is already up-to-date!
[nltk_data]    | Downloading package state_union to /root/nltk_data…
[nltk_data]    |   Package state_union is already up-to-date!
[nltk_data]    | Downloading package stopwords to /root/nltk_data…
[nltk_data]    |   Package stopwords is already up-to-date!
[nltk_data]    | Downloading package subjectivity to
[nltk_data]    |     /root/nltk_data…
[nltk_data]    |   Package subjectivity is already up-to-date!
[nltk_data]    | Downloading package swadesh to /root/nltk_data…
[nltk_data]    |   Package swadesh is already up-to-date!
[nltk_data]    | Downloading package switchboard to /root/nltk_data…
[nltk_data]    |   Package switchboard is already up-to-date!
[nltk_data]    | Downloading package tagsets to /root/nltk_data…
[nltk_data]    |   Package tagsets is already up-to-date!
```

```
[nltk_data]     | Downloading package tagsets_json to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package tagsets_json is already up-to-date!
[nltk_data]     | Downloading package timit to /root/nltk_data…
[nltk_data]     |   Package timit is already up-to-date!
[nltk_data]     | Downloading package toolbox to /root/nltk_data…
[nltk_data]     |   Package toolbox is already up-to-date!
[nltk_data]     | Downloading package treebank to /root/nltk_data…
[nltk_data]     |   Package treebank is already up-to-date!
[nltk_data]     | Downloading package twitter_samples to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package twitter_samples is already up-to-date!
[nltk_data]     | Downloading package udhr to /root/nltk_data…
[nltk_data]     |   Package udhr is already up-to-date!
[nltk_data]     | Downloading package udhr2 to /root/nltk_data…
[nltk_data]     |   Package udhr2 is already up-to-date!
[nltk_data]     | Downloading package unicode_samples to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package unicode_samples is already up-to-date!
[nltk_data]     | Downloading package universal_tagset to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package universal_tagset is already up-to-date!
[nltk_data]     | Downloading package universal_treebanks_v20 to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package universal_treebanks_v20 is already up-to-
[nltk_data]     |       date!
[nltk_data]     | Downloading package vader_lexicon to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package vader_lexicon is already up-to-date!
[nltk_data]     | Downloading package verbnet to /root/nltk_data…
[nltk_data]     |   Package verbnet is already up-to-date!
[nltk_data]     | Downloading package verbnet3 to /root/nltk_data…
[nltk_data]     |   Package verbnet3 is already up-to-date!
[nltk_data]     | Downloading package webtext to /root/nltk_data…
[nltk_data]     |   Package webtext is already up-to-date!
[nltk_data]     | Downloading package wmt15_eval to /root/nltk_data…
[nltk_data]     |   Package wmt15_eval is already up-to-date!
[nltk_data]     | Downloading package word2vec_sample to
[nltk_data]     |     /root/nltk_data…
[nltk_data]     |   Package word2vec_sample is already up-to-date!
[nltk_data]     | Downloading package wordnet to /root/nltk_data…
[nltk_data]     |   Package wordnet is already up-to-date!
[nltk_data]     | Downloading package wordnet2021 to /root/nltk_data…
[nltk_data]     |   Package wordnet2021 is already up-to-date!
[nltk_data]     | Downloading package wordnet2022 to /root/nltk_data…
[nltk_data]     |   Package wordnet2022 is already up-to-date!
[nltk_data]     | Downloading package wordnet31 to /root/nltk_data…
[nltk_data]     |   Package wordnet31 is already up-to-date!
```

```
[nltk_data]     | Downloading package wordnet_ic to /root/nltk_data…
[nltk_data]     |   Package wordnet_ic is already up-to-date!
[nltk_data]     | Downloading package words to /root/nltk_data…
[nltk_data]     |   Package words is already up-to-date!
[nltk_data]     | Downloading package ycoe to /root/nltk_data…
[nltk_data]     |   Package ycoe is already up-to-date!
[nltk_data]     |
[nltk_data]   Done downloading collection all
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[8]: True

1 Task 01: Preparing Data

[4]:
```
# Defining the file paths and their corresponding categories
file_paths = {
    "/content/drive/MyDrive/Data/Data/Business.xlsx": "Business",
    "/content/drive/MyDrive/Data/Data/Opinion.xlsx": "Opinion",
    "/content/drive/MyDrive/Data/Data/Political_gossip.xlsx": "Political␣
 ↪Gossip",
    "/content/drive/MyDrive/Data/Data/Sports.xlsx": "Sports",
    "/content/drive/MyDrive/Data/Data/World_news.xlsx": "World News"
}

# Initialize an empty list to store DataFrames
dfs = []

# Read each file, add the 'class' column, and drop 'title'
for file, category in file_paths.items():
    df = pd.read_excel(file)  # Read Excel file
    df["class"] = category  # Add class column
    df.drop(columns=["title"], inplace=True)  # Drop 'title' column
    dfs.append(df)  # Append to list

# Merge all DataFrames into a single DataFrame
final_df = pd.concat(dfs, ignore_index=True)

# Check for missing values
print("\nMissing values before treatment:")
print(final_df.isnull().sum())
```

```python
# Handling missing values
final_df["class"].fillna("Unknown", inplace=True)  # Fill missing classes if any
final_df.dropna(subset=["content"], inplace=True)  # Remove rows with missing␣
 ↪content

# Check again after treatment
print("\nMissing values after treatment:")
print(final_df.isnull().sum())

# Remove duplicates
final_df.drop_duplicates(inplace=True)

# Handle missing values (replace NaN with empty strings)
df["content"].fillna("", inplace=True)

# Save the final dataset as an Excel file
final_df.to_excel("Daily_Mirror_News.xlsx", index=False)
print("\nDataset saved as 'Daily_Mirror_News.xlsx'")
```

```
Missing values before treatment:
Unnamed: 0    0
content       4
class         0
dtype: int64

Missing values after treatment:
Unnamed: 0    0
content       0
class         0
dtype: int64
<ipython-input-4-c0c7c66b2c7a>:28: FutureWarning: A value is trying to be set on
a copy of a DataFrame or Series through chained assignment using an inplace
method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves as
a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  final_df["class"].fillna("Unknown", inplace=True)  # Fill missing classes if
any
<ipython-input-4-c0c7c66b2c7a>:39: FutureWarning: A value is trying to be set on
a copy of a DataFrame or Series through chained assignment using an inplace
```

```
method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves as
a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  df["content"].fillna("", inplace=True)


Dataset saved as 'Daily_Mirror_News.xlsx'
```

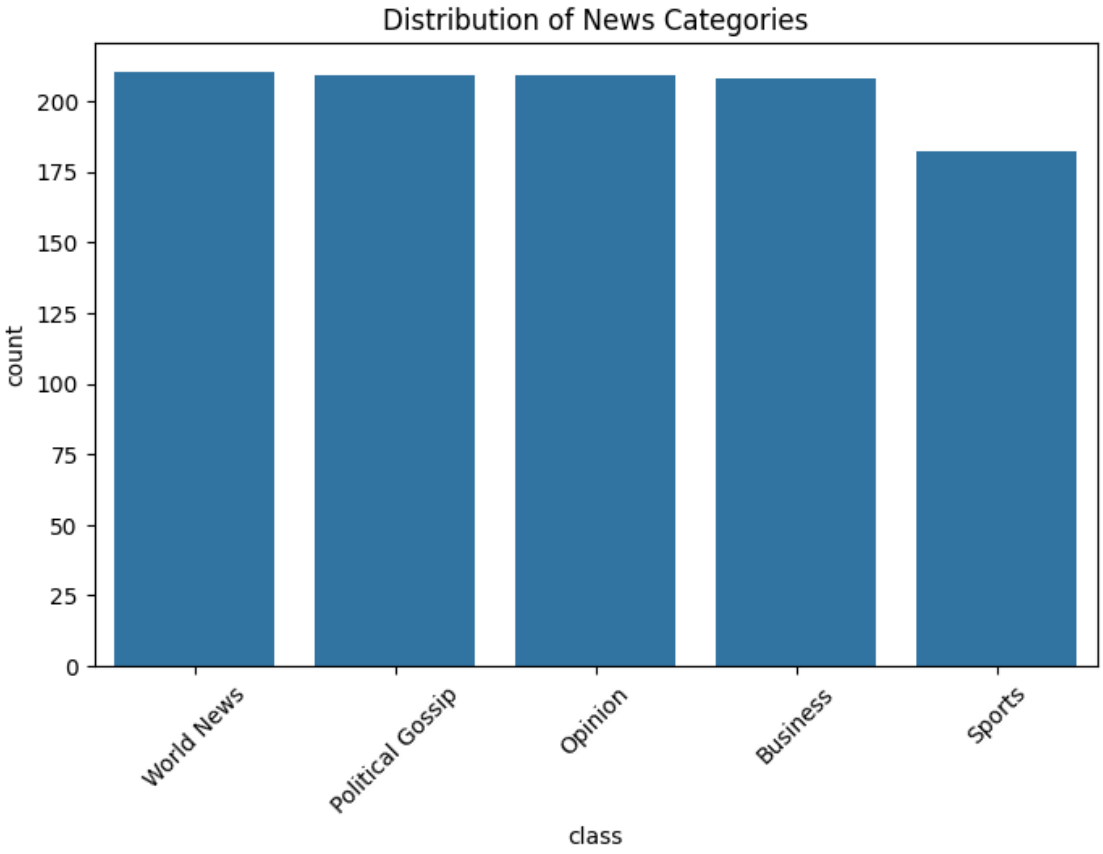
2 Task 2: EDA & Text Preprocessing

[9]:
```
# ===== TASK 2: EDA =====

# Load cleaned dataset
df = pd.read_excel("Daily_Mirror_News.xlsx")

# Word Cloud
text_data = " ".join(df["content"])

wordcloud = WordCloud(width=800, height=400, background_color='white').
 ↪generate(text_data)

plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.title("Word Cloud of News Articles")
plt.show()

# N-grams Analysis (Bigrams)
vectorizer_bigram = CountVectorizer(ngram_range=(2,2), stop_words='english')
bigrams = vectorizer_bigram.fit_transform(df["content"])
bigram_counts = Counter(vectorizer_bigram.get_feature_names_out())

print("Most common bigrams:", bigram_counts.most_common(10))

# N-grams Analysis (Trigrams)
vectorizer_trigram = CountVectorizer(ngram_range=(3,3), stop_words='english')
trigrams = vectorizer_trigram.fit_transform(df["content"])
trigram_counts = Counter(vectorizer_trigram.get_feature_names_out())

print("Most common trigrams:", trigram_counts.most_common(10))
```

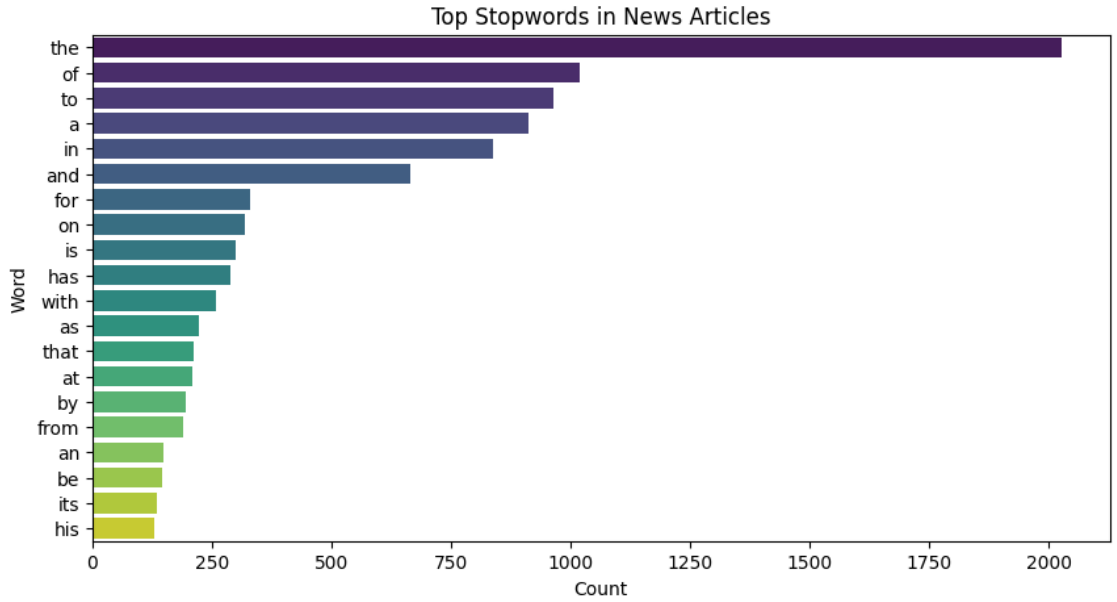
```
# Plot Class Distribution
plt.figure(figsize=(8,5))
sns.countplot(data=df, x='class', order=df['class'].value_counts().index)
plt.xticks(rotation=45)
plt.title("Distribution of News Categories")
plt.show()

# Analyze Stopwords (Find Most Frequent Stopwords)
stopwords = set(STOPWORDS)
stopword_counts = Counter([word for word in text_data.split() if word in␣
 ↪stopwords])

# Convert to DataFrame and plot
stopwords_df = pd.DataFrame(stopword_counts.items(), columns=["Word", "Count"]).
 ↪sort_values(by="Count", ascending=False)

plt.figure(figsize=(10,5))
sns.barplot(data=stopwords_df.head(20), x="Count", y="Word", palette="viridis")
plt.title("Top Stopwords in News Articles")
plt.show()
```


```
Most common bigrams: [('000 cash', 1), ('000 guests', 1), ('000 mark', 1), ('000
month', 1), ('000 new', 1), ('000 owed', 1), ('000 people', 1), ('000
purchased', 1), ('000 ramadan', 1), ('000 rs', 1)]
Most common trigrams: [('000 cash guilty', 1), ('000 mark days', 1), ('000 mark
hitting', 1), ('000 month cost', 1), ('000 new entrepreneurs', 1), ('000 owed
```

```
ific', 1), ('000 people evacuated', 1), ('000 people killed', 1), ('000
purchased pharmacy', 1), ('000 rs 60', 1)]
```




```
<ipython-input-9-90bae565303b>:46: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(data=stopwords_df.head(20), x="Count", y="Word",
palette="viridis")
```



[26]:
```
# ===== TASK 2: Text Preprocessing Steps =====

# Load the dataset
df = pd.read_excel("Daily_Mirror_News.xlsx")

# Selecting a sample text
sample_text = df["content"].iloc[0]  # Choosing row 1

# Define preprocessing function
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()

def preprocess_text_step_by_step(text, is_sample=False):
    # Step 1: Convert to Lowercase - Ensures uniformity, avoids duplication of␣
 ↪tokens, and simplifies further processing.
    text_step1 = text.lower()
    if is_sample:
        print("\n Step 1: Convert to Lowercase")
        print("Before:", text)
        print("After:", text_step1)

    # Step 2: Remove Special Characters & Numbers - Reduces noise and focuses␣
 ↪the analysis on meaningful words.
    text_step2 = re.sub(r'[^a-z\s]', '', text_step1)
    if is_sample:
        print("\n Step 2: Remove Special Characters & Numbers")
        print("Before:", text_step1)
```

```
        print("After:", text_step2)

    # Step 3: Tokenization - Breaks down the text into manageable units␣
 ↪(tokens) for further analysis.
    tokens = word_tokenize(text_step2)
    if is_sample:
        print("\n Step 3: Tokenization")
        print("Before:", text_step2)
        print("After:", tokens)

    # Step 4: Remove Stopwords - Eliminates frequent, meaningless words that do␣
 ↪not contribute to the overall meaning of the text.
    tokens_step4 = [word for word in tokens if word not in stop_words]
    if is_sample:
        print("\n Step 4: Remove Stopwords")
        print("Before:", tokens)
        print("After:", tokens_step4)

    # Step 5: Lemmatization - Normalizes words to their base form, reducing␣
 ↪variations and improving consistency in analysis.
    tokens_step5 = [lemmatizer.lemmatize(word) for word in tokens_step4]
    if is_sample:
        print("\n Step 5: Lemmatization")
        print("Before:", tokens_step4)
        print("After:", tokens_step5)

    return " ".join(tokens_step5)

# Apply preprocessing for the entire dataset
df["processed_content"] = df["content"].apply(lambda x:␣
 ↪preprocess_text_step_by_step(x, is_sample=False))

# Print the detailed steps for the first article
print("\n Detailed Steps for the First Article:")
preprocess_text_step_by_step(sample_text, is_sample=True)

# Save preprocessed dataset as 'Preprocessed_Daily_Mirror_News.xlsx'
df[["processed_content", "class"]].to_excel("Preprocessed_Daily_Mirror_News.
 ↪xlsx", index=False)

print("\n Preprocessing completed and saved as 'Preprocessed_Daily_Mirror_News.
 ↪xlsx'")
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
```

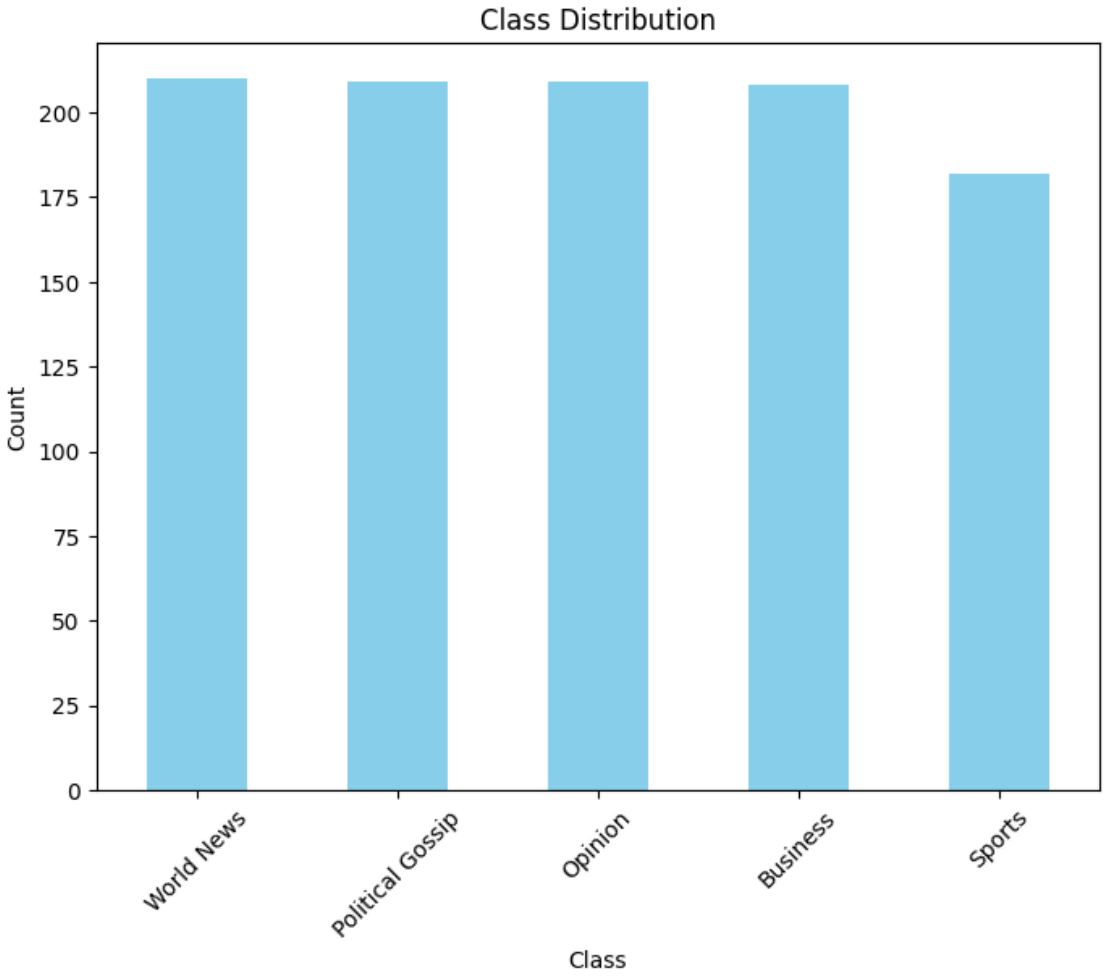
```
[nltk_data] Downloading package wordnet to /root/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!


 Detailed Steps for the First Article:

 Step 1: Convert to Lowercase
Before: Sri Lanka’s inflation is expected to increase "sizeably" in the third
quarter this year (3Q25), with the possibility of it hovering around 2
percentage points above the inflation target in mid-2026, the Central Bank of
Sri Lanka (CBSL) said in its monetary policy report that was released yesterday.
After: sri lanka’s inflation is expected to increase "sizeably" in the third
quarter this year (3q25), with the possibility of it hovering around 2
percentage points above the inflation target in mid-2026, the central bank of
sri lanka (cbsl) said in its monetary policy report that was released yesterday.

 Step 2: Remove Special Characters & Numbers
Before: sri lanka’s inflation is expected to increase "sizeably" in the third
quarter this year (3q25), with the possibility of it hovering around 2
percentage points above the inflation target in mid-2026, the central bank of
sri lanka (cbsl) said in its monetary policy report that was released yesterday.
After: sri lankas inflation is expected to increase sizeably in the third
quarter this year q with the possibility of it hovering around  percentage
points above the inflation target in mid the central bank of sri lanka cbsl said
in its monetary policy report that was released yesterday

 Step 3: Tokenization
Before: sri lankas inflation is expected to increase sizeably in the third
quarter this year q with the possibility of it hovering around  percentage
points above the inflation target in mid the central bank of sri lanka cbsl said
in its monetary policy report that was released yesterday
After: ['sri', 'lankas', 'inflation', 'is', 'expected', 'to', 'increase',
'sizeably', 'in', 'the', 'third', 'quarter', 'this', 'year', 'q', 'with', 'the',
'possibility', 'of', 'it', 'hovering', 'around', 'percentage', 'points',
'above', 'the', 'inflation', 'target', 'in', 'mid', 'the', 'central', 'bank',
'of', 'sri', 'lanka', 'cbsl', 'said', 'in', 'its', 'monetary', 'policy',
'report', 'that', 'was', 'released', 'yesterday']

 Step 4: Remove Stopwords
Before: ['sri', 'lankas', 'inflation', 'is', 'expected', 'to', 'increase',
'sizeably', 'in', 'the', 'third', 'quarter', 'this', 'year', 'q', 'with', 'the',
'possibility', 'of', 'it', 'hovering', 'around', 'percentage', 'points',
'above', 'the', 'inflation', 'target', 'in', 'mid', 'the', 'central', 'bank',
'of', 'sri', 'lanka', 'cbsl', 'said', 'in', 'its', 'monetary', 'policy',
'report', 'that', 'was', 'released', 'yesterday']
After: ['sri', 'lankas', 'inflation', 'expected', 'increase', 'sizeably',
'third', 'quarter', 'year', 'q', 'possibility', 'hovering', 'around',
'percentage', 'points', 'inflation', 'target', 'mid', 'central', 'bank', 'sri',
```

```
'lanka', 'cbsl', 'said', 'monetary', 'policy', 'report', 'released',
'yesterday']

 Step 5: Lemmatization
Before: ['sri', 'lankas', 'inflation', 'expected', 'increase', 'sizeably',
'third', 'quarter', 'year', 'q', 'possibility', 'hovering', 'around',
'percentage', 'points', 'inflation', 'target', 'mid', 'central', 'bank', 'sri',
'lanka', 'cbsl', 'said', 'monetary', 'policy', 'report', 'released',
'yesterday']
After: ['sri', 'lankas', 'inflation', 'expected', 'increase', 'sizeably',
'third', 'quarter', 'year', 'q', 'possibility', 'hovering', 'around',
'percentage', 'point', 'inflation', 'target', 'mid', 'central', 'bank', 'sri',
'lanka', 'cbsl', 'said', 'monetary', 'policy', 'report', 'released',
'yesterday']

 Preprocessing completed and saved as 'Preprocessed_Daily_Mirror_News.xlsx'
```

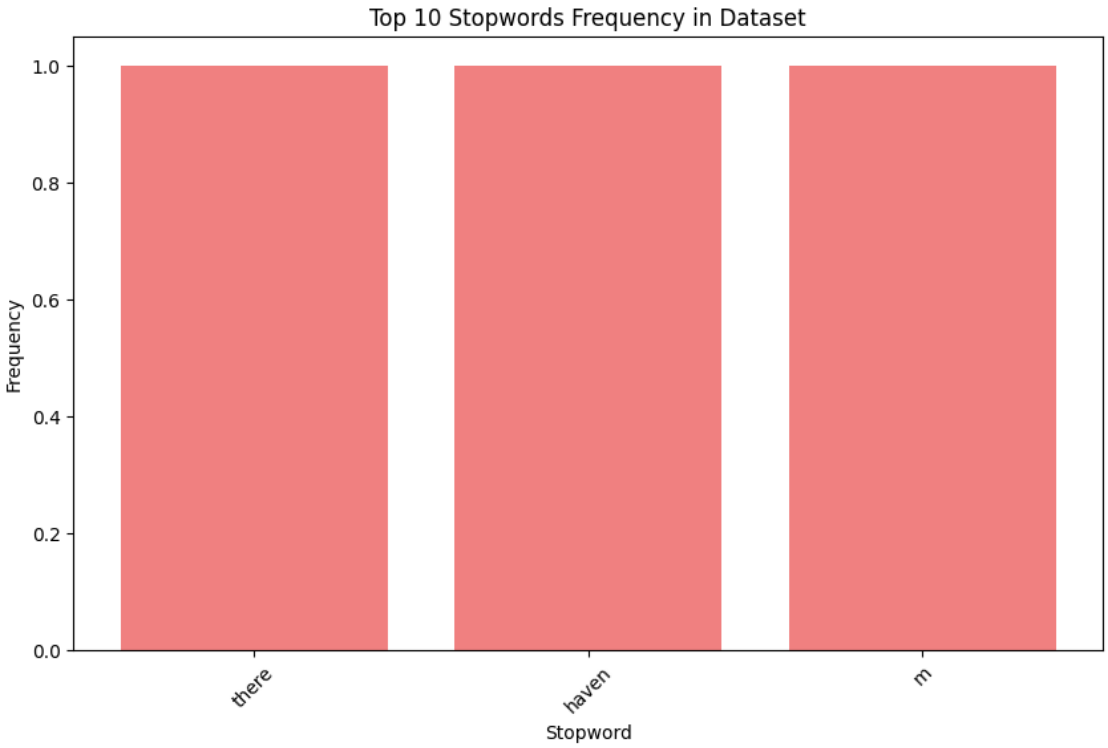
[27]:
```
# EDA after preprocessing

# 1. Word Cloud After Preprocessing
text_data_cleaned = " ".join(df["processed_content"])

wordcloud_cleaned = WordCloud(width=800, height=400, background_color='white').
 ↪generate(text_data_cleaned)

plt.figure(figsize=(10, 5))
plt.imshow(wordcloud_cleaned, interpolation='bilinear')
plt.axis("off")
plt.title("Word Cloud After Preprocessing")
plt.show()

# 2. N-Grams After Preprocessing (Bigram Example)
vectorizer = CountVectorizer(ngram_range=(2, 2), stop_words='english')
bigrams_cleaned = vectorizer.fit_transform(df["processed_content"])
bigram_counts_cleaned = Counter(vectorizer.get_feature_names_out())

print("Most common bigrams after preprocessing:", bigram_counts_cleaned.
 ↪most_common(10))

# 3. Plot Class Distribution
class_counts = df["class"].value_counts()

plt.figure(figsize=(8, 6))
class_counts.plot(kind='bar', color='skyblue')
plt.title('Class Distribution')
plt.xlabel('Class')
plt.ylabel('Count')
```

```
plt.xticks(rotation=45)
plt.show()

# 4. Stopwords EDA
# Load the stopwords from NLTK
stop_words = set(stopwords.words('english'))

# Calculate stopwords frequency in the entire dataset
stopwords_in_data = [word for content in df["processed_content"] for word in␣
 ↪content.split() if word in stop_words]

# Count stopwords frequency
stopword_counts = Counter(stopwords_in_data)

# Display the most common stopwords
print("\nMost Common Stopwords in the Dataset:")
for word, count in stopword_counts.most_common(10):
    print(f"{word}: {count}")

# Plot the distribution of stopwords in the dataset
plt.figure(figsize=(10, 6))
plt.bar([word for word, _ in stopword_counts.most_common(10)], [count for _,␣
 ↪count in stopword_counts.most_common(10)], color='lightcoral')
plt.title('Top 10 Stopwords Frequency in Dataset')
plt.xlabel('Stopword')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.show()
```

```
Most common bigrams after preprocessing: [('aac introduce', 1), ('aalka aalka',
1), ('aalka stable', 1), ('ab mauri', 1), ('abans plc', 1), ('abc news', 1),
('abducting driver', 1), ('abeywardana said', 1), ('ability recover', 1), ('able
muster', 1)]
```




```
Most Common Stopwords in the Dataset:
there: 1
haven: 1
m: 1
```



3 Task 3: Select a Hugging Face Model

For this text classification task involving news articles categorized into multiple classes such as Business, Opinion, Political Gossip, World News, and Evaluation, the model selected from Hugging Face is distilbert-base-uncased. This model is a distilled, lightweight version of the original BERT (bert-base-uncased) and retains over 95% of BERT's performance while being significantly faster and more efficient in terms of computation and memory usage. Since the objective is to build a classifier that assigns one of several labels to each news article, distilbert-base-uncased is an appropriate choice as it is pretrained on large English corpora and effectively captures contextual representations of words. It can be fine-tuned using the DistilBertForSequenceClassification architecture by specifying the number of output labels (in this case, five). This model offers an optimal balance between accuracy and resource efficiency, making it well-suited for fine-tuning on the current dataset within environments that may have limited computational capacity.

4 Task 4: Finetune a Hugging Face Model

[ ]:
```
# Step 1:Importing libraries
import evaluate
import numpy as np
import pandas as pd

from IPython.display import clear_output
```

```
from huggingface_hub import notebook_login
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from datasets import Dataset, DatasetDict
from transformers import DistilBertForSequenceClassification
from transformers import TrainingArguments
from transformers import Trainer
from transformers import AutoTokenizer
from datasets import Dataset, DatasetDict
from transformers import EarlyStoppingCallback
from sklearn.metrics import precision_recall_fscore_support
```

[ ]:
```
# Step 2:Load Preprocessed Dataset
data = pd.read_excel("Preprocessed_Daily_Mirror_News.xlsx")
```

[ ]:
```
# Step 3:Conversion of categorical 'class' values to numerical labels

# Import LabelEncoder from sklearn
from sklearn.preprocessing import LabelEncoder

# Create a LabelEncoder object
label_encoder = LabelEncoder()

# Encode the 'class' column with numerical labels (0, 1, 2,...)
data['label'] = label_encoder.fit_transform(data['class'])

# Print the unique values of the encoded 'label' column
print(data['label'].unique())

# Display the fitted classes (i.e., the original labels that correspond to each␣
 ↪numerical value)
print(label_encoder.classes_)
```

```
[0 1 2 3 4]
['Business' 'Opinion' 'Political Gossip' 'Sports' 'World News']
```

[ ]:
```
# Step 4:Loading the tokenizer and model

from transformers import AutoTokenizer, AutoModelForSequenceClassification

# Selected Model: DistilBERT
model_checkpoint = "distilbert-base-uncased"

# Justification:
"""
We chose 'distilbert-base-uncased' as the base model due to the following␣
 ↪reasons:
```

```
1. It is a lightweight version of BERT and is significantly faster while␣
 ↪maintaining performance.
2. Ideal for fine-tuning on relatively small datasets like ours.
3. It is widely used and well-supported by Hugging Face for text classification␣
 ↪tasks.
4. Lower computational cost makes it suitable for hosting in WebApps␣
 ↪(especially Hugging Face Spaces).
"""

# Load tokenizer and base model for inspection
tokenizer = AutoTokenizer.from_pretrained(model_checkpoint)
model = AutoModelForSequenceClassification.from_pretrained(model_checkpoint,␣
 ↪num_labels=5)

print(f"Model '{model_checkpoint}' loaded with {model.num_labels} labels for␣
 ↪fine-tuning.")
```

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

```
Model 'distilbert-base-uncased' loaded with 5 labels for fine-tuning.
```

[ ]:
```
# Step 5: Split into Train & Validation (80%-20%)
train_texts, val_texts, train_labels, val_labels = train_test_split(
    data["processed_content"], data["label"], test_size=0.2, random_state=42
)

# Step 6: Load Tokenizer & Model
model_checkpoint = "distilbert-base-uncased"
tokenizer = AutoTokenizer.from_pretrained(model_checkpoint)
model = DistilBertForSequenceClassification.from_pretrained(model_checkpoint,␣
 ↪num_labels=5)

# Step 7: Tokenize the Text
def tokenize_function(examples):
    return tokenizer(examples["text"], padding="max_length", truncation=True,␣
 ↪max_length=512)

tokenized_train_dataset = train_dataset.map(tokenize_function, batched=True)
tokenized_val_dataset = val_dataset.map(tokenize_function, batched=True)

# Remove original text column (only keep tokenized features)
tokenized_train_dataset = tokenized_train_dataset.remove_columns(["text"])
```

```python
tokenized_val_dataset = tokenized_val_dataset.remove_columns(["text"])

# Convert into a DatasetDict
hf_dataset = DatasetDict({"train": tokenized_train_dataset, "validation":␣
↪tokenized_val_dataset})

# Step 8: Check the tokenized datasets
print(tokenized_train_dataset[5]['input_ids'], '\n')
print(tokenized_train_dataset[5]['attention_mask'], '\n')
```

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

```
Map:   0%|          | 0/814 [00:00<?, ? examples/s]
```

```
Map:   0%|          | 0/204 [00:00<?, ? examples/s]
```

```
[101, 2533, 6870, 5680, 2405, 3343, 4861, 3780, 6224, 2270, 5448, 2144, 2146,
3334, 2213, 6378, 3213, 7864, 2533, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

```
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

[ ]:
```python
from evaluate import load

# Step 9: Define Training Arguments
args = TrainingArguments(
    output_dir="HuggingFaceAttempt1",
    run_name="version1",
    evaluation_strategy="steps",  # Evaluate after every few steps
    eval_steps=100,  # Evaluate after every 100 steps
    per_device_train_batch_size=4,  # Training batch size
    per_device_eval_batch_size=4,  # Evaluation batch size
    num_train_epochs=3,  # Number of epochs
    seed=0,  # Set the random seed
    load_best_model_at_end=True,  # Pick the best model based on validation
    report_to=None,  # Avoid using W&B by not reporting to it
    save_steps=500,  # Save model every 10 steps

)

# Step 10: Define Metrics
metric = load("accuracy")

def compute_metrics(p):
    '''
    This function calculates accuracy, precision, recall, and f1 scores by␣
 ↪comparing
    predicted values with true (reference) values.
    '''
    predictions, references = p
    preds = np.argmax(predictions, axis=1)  # Convert logits to class labels
    # Compute all relevant metrics (accuracy, precision, recall, and f1)
    return {
        "accuracy": metric.compute(predictions=preds, references=references),
```

```
        "precision": precision_recall_fscore_support(references, preds,␣
 ↪average="macro")[0],
        "recall": precision_recall_fscore_support(references, preds,␣
 ↪average="macro")[1],
        "f1": precision_recall_fscore_support(references, preds,␣
 ↪average="macro")[2],
    }

# Step 11: Setup Trainer
trainer = Trainer(
    model=model,
    args=args,
    train_dataset=hf_dataset["train"],
    eval_dataset=hf_dataset["validation"],
    compute_metrics=compute_metrics,
    callbacks=[EarlyStoppingCallback(early_stopping_patience=3)],
)

# Set the model's label to class mapping
model.config.id2label = {
    0: "Business",
    1: "Opinion",
    2: "Political_gossip",
    3: "Sports",
    4: "World_news"
}

# Create the reverse mapping from label to id
model.config.label2id = {v: k for k, v in model.config.id2label.items()}
```

```
Using the `WANDB_DISABLED` environment variable is deprecated and will be
removed in v5. Use the --report_to flag to control the integrations used for
logging result (for instance --report_to none).
```

[ ]:
```
# Disable W&B login by setting the environment variable early
import os
os.environ["WANDB_DISABLED"] = "true"  # Disable W&B
import wandb
```

Fine-tuning the Model:

[ ]:
```
# Step 12: Start Fine-Tuning
train_output = trainer.train()

# Save trained model
trainer.save_model("content/")
```

```
# Print the train output
print(train_output)

print(" Fine-tuning complete!")
```

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

Trainer is attempting to log a value of "{'accuracy': 0.8676470588235294}" of
type <class 'dict'> for key "eval/accuracy" as a scalar. This invocation of
Tensorboard's writer.add_scalar() is incorrect so we dropped this attribute.
Trainer is attempting to log a value of "{'accuracy': 0.8823529411764706}" of
type <class 'dict'> for key "eval/accuracy" as a scalar. This invocation of
Tensorboard's writer.add_scalar() is incorrect so we dropped this attribute.
Trainer is attempting to log a value of "{'accuracy': 0.8872549019607843}" of
type <class 'dict'> for key "eval/accuracy" as a scalar. This invocation of
Tensorboard's writer.add_scalar() is incorrect so we dropped this attribute.
Trainer is attempting to log a value of "{'accuracy': 0.8872549019607843}" of
type <class 'dict'> for key "eval/accuracy" as a scalar. This invocation of
Tensorboard's writer.add_scalar() is incorrect so we dropped this attribute.
Trainer is attempting to log a value of "{'accuracy': 0.9019607843137255}" of
type <class 'dict'> for key "eval/accuracy" as a scalar. This invocation of
Tensorboard's writer.add_scalar() is incorrect so we dropped this attribute.

TrainOutput(global_step=500, training_loss=0.3633096008300781,
metrics={'train_runtime': 8260.8225, 'train_samples_per_second': 0.296,
'train_steps_per_second': 0.074, 'total_flos': 264419073576960.0, 'train_loss':
0.3633096008300781, 'epoch': 2.450980392156863})
 Fine-tuning complete!
```

Saving the Model:

[ ]:
```
# Log into Hugging Face (fixes 401 Unauthorized issue)
from huggingface_hub import notebook_login
notebook_login()

# Save model and tokenizer locally
model.save_pretrained("News_Classification_Model")
tokenizer.save_pretrained("News_Classification_Model")  # Save tokenizer locally
```

```
VBox(children=(HTML(value='<center> <img\nsrc=https://huggingface.co/front/
↪assets/huggingface_logo-noborder.sv…
```

[ ]:
```
('News_Classification_Model/tokenizer_config.json',
 'News_Classification_Model/special_tokens_map.json',
 'News_Classification_Model/vocab.txt',
 'News_Classification_Model/added_tokens.json',
 'News_Classification_Model/tokenizer.json')
```

[ ]:
```
# Pushes the Model and Tokenizer to Hugging Face Hub
model.push_to_hub("TAgroup5/news-classification-model")
tokenizer.push_to_hub("TAgroup5/news-classification-model")
```

```
model.safetensors:   0%|          | 0.00/268M [00:00<?, ?B/s]
```

```
No files have been modified since last commit. Skipping to prevent empty commit.
WARNING:huggingface_hub.hf_api:No files have been modified since last commit.
Skipping to prevent empty commit.
```

[ ]:
```
CommitInfo(commit_url='https://huggingface.co/TAgroup5/news-classification-
model/commit/2cb540dc84e4c1e32677882b59b4114dbfb44b11', commit_message='Upload
tokenizer', commit_description='',
oid='2cb540dc84e4c1e32677882b59b4114dbfb44b11', pr_url=None,
repo_url=RepoUrl('https://huggingface.co/TAgroup5/news-classification-model',
endpoint='https://huggingface.co', repo_type='model', repo_id='TAgroup5/news-
classification-model'), pr_revision=None, pr_num=None)
```

[ ]:
```
# Loads the model
new_model = "TAgroup5/news-classification-model"
```

Calling the Pre-trained Model:

[ ]:
```
# Load model and tokenizer from Hugging Face Hub
from transformers import pipeline

# Use Hugging Face pipeline to classify the text
pipe = pipeline("text-classification", model="TAgroup5/
 ↪news-classification-model")

text = '''
Emerging threats in maritime waters have posed significant challenges to the␣
 ↪land-locked as well as littoral states such as Sri Lanka.'''
# Get prediction
predictions = pipe(text)
print(predictions)
```

```
Device set to use cpu
```

```
[{'label': 'Opinion', 'score': 0.9944149255752563}]
```

[28]:
```
# Include the link to your fine-tuned model pushed to Hugging Face
model_link = "https://huggingface.co/spaces/TAgroup5/demo-News_classifier"
print("Fine-tuned model is available at:", model_link)
```

```
Fine-tuned model is available at: https://huggingface.co/spaces/TAgroup5/demo-
News_classifier
```