HOMEWORK 6: TEXT CLASSIFICATION

In this homework, you will create models to classify texts from TRUE call-center. There are two classification tasks:

1. Action Classification: Identify which action the customer would like to take (e.g. enquire, report, cancle)
2. Object Classification: Identify which object the customer is referring to (e.g. payment, truemoney, internet, roaming)

We will focus only on the Object Classification task for this homework.

In this homework, you are asked compare different text classification models in terms of accuracy and inference time.

You will need to build 3 different models.

1. A model based on tf-idf
2. A model based on MUSE
3. A model based on wangchanBERTa

You will be ask to submit 3 different files (.pdf from .ipynb) that does the 3 different models. Finally, answer the accuracy and runtime numbers in MCV.

This homework is quite free form, and your answer may vary. We hope that the processing during the course of this assignment will make you think more about the design choices in text classification.

```
!wget --no-check-certificate
https://www.dropbox.com/s/37u83g55p19kvrl/clean-phone-data-for-
students.csv

--2025-02-15 17:25:49--
https://www.dropbox.com/s/37u83g55p19kvrl/clean-phone-data-for-
students.csv
Resolving www.dropbox.com (www.dropbox.com)... 162.125.3.18,
2620:100:6018:18::a27d:312
Connecting to www.dropbox.com (www.dropbox.com)|162.125.3.18|:443...
connected.
HTTP request sent, awaiting response... 302 Found
Location: https://www.dropbox.com/scl/fi/8h8hvsw9uj6o0524lfe4i/clean-
phone-data-for-students.csv?rlkey=lwv5xbf16jerehnv3lfgq5ue6
[following]
--2025-02-15 17:25:49--
https://www.dropbox.com/scl/fi/8h8hvsw9uj6o0524lfe4i/clean-phone-data-
for-students.csv?rlkey=lwv5xbf16jerehnv3lfgq5ue6
Reusing existing connection to www.dropbox.com:443.
HTTP request sent, awaiting response... 302 Found
```

```
Location:
https://ucd140658d06553ba0fb5f39edc4.dl.dropboxusercontent.com/cd/0/
inline/CkKTYFCxdlz7-
k1p7pp6ebl7mdHz3qc9Me3wCOfH7Qpj7N2tzFBWDmdb9ghn5fc5B0bYSN7Tl29IaoqMKf8
YBtC4PAviOqFCeB2KcTZJkgOpMXJqVbbmbPP6Dt2s1xG4MKs/file# [following]
--2025-02-15 17:25:49--
https://ucd140658d06553ba0fb5f39edc4.dl.dropboxusercontent.com/cd/0/
inline/CkKTYFCxdlz7-
k1p7pp6ebl7mdHz3qc9Me3wCOfH7Qpj7N2tzFBWDmdb9ghn5fc5B0bYSN7Tl29IaoqMKf8
YBtC4PAviOqFCeB2KcTZJkgOpMXJqVbbmbPP6Dt2s1xG4MKs/file
Resolving ucd140658d06553ba0fb5f39edc4.dl.dropboxusercontent.com
(ucd140658d06553ba0fb5f39edc4.dl.dropboxusercontent.com)...
162.125.3.15, 2620:100:6018:15::a27d:30f
Connecting to ucd140658d06553ba0fb5f39edc4.dl.dropboxusercontent.com
(ucd140658d06553ba0fb5f39edc4.dl.dropboxusercontent.com)|
162.125.3.15|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2518977 (2.4M) [text/plain]
Saving to: ‘clean-phone-data-for-students.csv.1’

clean-phone-data-fo 100%[===================>]   2.40M  --.-KB/s    in
0.08s

2025-02-15 17:25:50 (30.4 MB/s) - ‘clean-phone-data-for-
students.csv.1’ saved [2518977/2518977]


!pip install pythainlp
!pip install -U sentence-transformers
!pip install tf-keras

Requirement already satisfied: pythainlp in
/usr/local/lib/python3.10/dist-packages (5.0.5)
Requirement already satisfied: requests>=2.22.0 in
/usr/local/lib/python3.10/dist-packages (from pythainlp) (2.32.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.22.0-
>pythainlp) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.22.0-
>pythainlp) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.22.0-
>pythainlp) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.22.0-
>pythainlp) (2025.1.31)
Requirement already satisfied: sentence-transformers in
/usr/local/lib/python3.10/dist-packages (3.4.1)
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in
```

```
/usr/local/lib/python3.10/dist-packages (from sentence-transformers)
(4.47.0)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-
packages (from sentence-transformers) (4.67.1)
Requirement already satisfied: torch>=1.11.0 in
/usr/local/lib/python3.10/dist-packages (from sentence-transformers)
(2.5.1+cu121)
Requirement already satisfied: scikit-learn in
/usr/local/lib/python3.10/dist-packages (from sentence-transformers)
(1.2.2)
Requirement already satisfied: scipy in
/usr/local/lib/python3.10/dist-packages (from sentence-transformers)
(1.13.1)
Requirement already satisfied: huggingface-hub>=0.20.0 in
/usr/local/lib/python3.10/dist-packages (from sentence-transformers)
(0.28.1)
Requirement already satisfied: Pillow in
/usr/local/lib/python3.10/dist-packages (from sentence-transformers)
(11.0.0)
Requirement already satisfied: filelock in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (3.17.0)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (2024.9.0)
Requirement already satisfied: packaging>=20.9 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (6.0.2)
Requirement already satisfied: requests in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (2.32.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (4.12.2)
Requirement already satisfied: networkx in
/usr/local/lib/python3.10/dist-packages (from torch>=1.11.0->sentence-
transformers) (3.4.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.11.0->sentence-
transformers) (3.1.4)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.11.0->sentence-
transformers) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from sympy==1.13.1-
>torch>=1.11.0->sentence-transformers) (1.3.0)
```

```
Requirement already satisfied: numpy>=1.17 in
/usr/local/lib/python3.10/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (1.26.4)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (2024.11.6)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/usr/local/lib/python3.10/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (0.21.0)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.10/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (0.4.5)
Requirement already satisfied: joblib>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn->sentence-
transformers) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn->sentence-
transformers) (3.5.0)
Requirement already satisfied: mkl_fft in
/usr/local/lib/python3.10/dist-packages (from numpy>=1.17-
>transformers<5.0.0,>=4.41.0->sentence-transformers) (1.3.8)
Requirement already satisfied: mkl_random in
/usr/local/lib/python3.10/dist-packages (from numpy>=1.17-
>transformers<5.0.0,>=4.41.0->sentence-transformers) (1.2.4)
Requirement already satisfied: mkl_umath in
/usr/local/lib/python3.10/dist-packages (from numpy>=1.17-
>transformers<5.0.0,>=4.41.0->sentence-transformers) (0.1.1)
Requirement already satisfied: mkl in /usr/local/lib/python3.10/dist-
packages (from numpy>=1.17->transformers<5.0.0,>=4.41.0->sentence-
transformers) (2025.0.1)
Requirement already satisfied: tbb4py in
/usr/local/lib/python3.10/dist-packages (from numpy>=1.17-
>transformers<5.0.0,>=4.41.0->sentence-transformers) (2022.0.0)
Requirement already satisfied: mkl-service in
/usr/local/lib/python3.10/dist-packages (from numpy>=1.17-
>transformers<5.0.0,>=4.41.0->sentence-transformers) (2.4.1)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->torch>=1.11.0-
>sentence-transformers) (3.0.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.20.0->sentence-transformers) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.20.0->sentence-transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.20.0->sentence-transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
```

```
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.20.0->sentence-transformers) (2025.1.31)
Requirement already satisfied: intel-openmp>=2024 in
/usr/local/lib/python3.10/dist-packages (from mkl->numpy>=1.17-
>transformers<5.0.0,>=4.41.0->sentence-transformers) (2024.2.0)
Requirement already satisfied: tbb==2022.* in
/usr/local/lib/python3.10/dist-packages (from mkl->numpy>=1.17-
>transformers<5.0.0,>=4.41.0->sentence-transformers) (2022.0.0)
Requirement already satisfied: tcmlib==1.* in
/usr/local/lib/python3.10/dist-packages (from tbb==2022.*->mkl-
>numpy>=1.17->transformers<5.0.0,>=4.41.0->sentence-transformers)
(1.2.0)
Requirement already satisfied: intel-cmplr-lib-rt in
/usr/local/lib/python3.10/dist-packages (from mkl_umath->numpy>=1.17-
>transformers<5.0.0,>=4.41.0->sentence-transformers) (2024.2.0)
Requirement already satisfied: intel-cmplr-lib-ur==2024.2.0 in
/usr/local/lib/python3.10/dist-packages (from intel-openmp>=2024->mkl-
>numpy>=1.17->transformers<5.0.0,>=4.41.0->sentence-transformers)
(2024.2.0)
Requirement already satisfied: tf-keras in
/usr/local/lib/python3.10/dist-packages (2.17.0)
Requirement already satisfied: tensorflow<2.18,>=2.17 in
/usr/local/lib/python3.10/dist-packages (from tf-keras) (2.17.1)
Requirement already satisfied: absl-py>=1.0.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (1.4.0)
Requirement already satisfied: astunparse>=1.6.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (1.6.3)
Requirement already satisfied: flatbuffers>=24.3.25 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (24.3.25)
Requirement already satisfied: gast!=0.5.0,!=0.5.1,!=0.5.2,>=0.2.1
in /usr/local/lib/python3.10/dist-packages (from
tensorflow<2.18,>=2.17->tf-keras) (0.6.0)
Requirement already satisfied: google-pasta>=0.1.1 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (0.2.0)
Requirement already satisfied: h5py>=3.10.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (3.12.1)
Requirement already satisfied: libclang>=13.0.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (18.1.1)
Requirement already satisfied: ml-dtypes<0.5.0,>=0.3.1 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (0.4.1)
Requirement already satisfied: opt-einsum>=2.3.2 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
```

```
>tf-keras) (3.4.0)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (24.2)
Requirement already satisfied: protobuf!=4.21.0,!=4.21.1,!=4.21.2,!
=4.21.3,!=4.21.4,!=4.21.5,<5.0.0dev,>=3.20.3 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (3.20.3)
Requirement already satisfied: requests<3,>=2.21.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (2.32.3)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (75.1.0)
Requirement already satisfied: six>=1.12.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (1.17.0)
Requirement already satisfied: termcolor>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (2.5.0)
Requirement already satisfied: typing-extensions>=3.6.6 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (4.12.2)
Requirement already satisfied: wrapt>=1.11.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (1.17.0)
Requirement already satisfied: grpcio<2.0,>=1.24.3 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (1.68.1)
Requirement already satisfied: tensorboard<2.18,>=2.17 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (2.17.1)
Requirement already satisfied: keras>=3.2.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (3.5.0)
Requirement already satisfied: tensorflow-io-gcs-filesystem>=0.23.1 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (0.37.1)
Requirement already satisfied: numpy<2.0.0,>=1.23.5 in
/usr/local/lib/python3.10/dist-packages (from tensorflow<2.18,>=2.17-
>tf-keras) (1.26.4)
Requirement already satisfied: wheel<1.0,>=0.23.0 in
/usr/local/lib/python3.10/dist-packages (from astunparse>=1.6.0-
>tensorflow<2.18,>=2.17->tf-keras) (0.45.1)
Requirement already satisfied: rich in /usr/local/lib/python3.10/dist-
packages (from keras>=3.2.0->tensorflow<2.18,>=2.17->tf-keras)
(13.9.4)
Requirement already satisfied: namex in
/usr/local/lib/python3.10/dist-packages (from keras>=3.2.0-
```

```
>tensorflow<2.18,>=2.17->tf-keras) (0.0.8)
Requirement already satisfied: optree in
/usr/local/lib/python3.10/dist-packages (from keras>=3.2.0-
>tensorflow<2.18,>=2.17->tf-keras) (0.13.1)
Requirement already satisfied: mkl_fft in
/usr/local/lib/python3.10/dist-packages (from numpy<2.0.0,>=1.23.5-
>tensorflow<2.18,>=2.17->tf-keras) (1.3.8)
Requirement already satisfied: mkl_random in
/usr/local/lib/python3.10/dist-packages (from numpy<2.0.0,>=1.23.5-
>tensorflow<2.18,>=2.17->tf-keras) (1.2.4)
Requirement already satisfied: mkl_umath in
/usr/local/lib/python3.10/dist-packages (from numpy<2.0.0,>=1.23.5-
>tensorflow<2.18,>=2.17->tf-keras) (0.1.1)
Requirement already satisfied: mkl in /usr/local/lib/python3.10/dist-
packages (from numpy<2.0.0,>=1.23.5->tensorflow<2.18,>=2.17->tf-keras)
(2025.0.1)
Requirement already satisfied: tbb4py in
/usr/local/lib/python3.10/dist-packages (from numpy<2.0.0,>=1.23.5-
>tensorflow<2.18,>=2.17->tf-keras) (2022.0.0)
Requirement already satisfied: mkl-service in
/usr/local/lib/python3.10/dist-packages (from numpy<2.0.0,>=1.23.5-
>tensorflow<2.18,>=2.17->tf-keras) (2.4.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2.21.0-
>tensorflow<2.18,>=2.17->tf-keras) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2.21.0-
>tensorflow<2.18,>=2.17->tf-keras) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2.21.0-
>tensorflow<2.18,>=2.17->tf-keras) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2.21.0-
>tensorflow<2.18,>=2.17->tf-keras) (2025.1.31)
Requirement already satisfied: markdown>=2.6.8 in
/usr/local/lib/python3.10/dist-packages (from tensorboard<2.18,>=2.17-
>tensorflow<2.18,>=2.17->tf-keras) (3.7)
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0
in /usr/local/lib/python3.10/dist-packages (from
tensorboard<2.18,>=2.17->tensorflow<2.18,>=2.17->tf-keras) (0.7.2)
Requirement already satisfied: werkzeug>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from tensorboard<2.18,>=2.17-
>tensorflow<2.18,>=2.17->tf-keras) (3.1.3)
Requirement already satisfied: MarkupSafe>=2.1.1 in
/usr/local/lib/python3.10/dist-packages (from werkzeug>=1.0.1-
>tensorboard<2.18,>=2.17->tensorflow<2.18,>=2.17->tf-keras) (3.0.2)
Requirement already satisfied: intel-openmp>=2024 in
/usr/local/lib/python3.10/dist-packages (from mkl-
>numpy<2.0.0,>=1.23.5->tensorflow<2.18,>=2.17->tf-keras) (2024.2.0)
```

```
Requirement already satisfied: tbb==2022.* in
/usr/local/lib/python3.10/dist-packages (from mkl-
>numpy<2.0.0,>=1.23.5->tensorflow<2.18,>=2.17->tf-keras) (2022.0.0)
Requirement already satisfied: tcmlib==1.* in
/usr/local/lib/python3.10/dist-packages (from tbb==2022.*->mkl-
>numpy<2.0.0,>=1.23.5->tensorflow<2.18,>=2.17->tf-keras) (1.2.0)
Requirement already satisfied: intel-cmplr-lib-rt in
/usr/local/lib/python3.10/dist-packages (from mkl_umath-
>numpy<2.0.0,>=1.23.5->tensorflow<2.18,>=2.17->tf-keras) (2024.2.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.10/dist-packages (from rich->keras>=3.2.0-
>tensorflow<2.18,>=2.17->tf-keras) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/usr/local/lib/python3.10/dist-packages (from rich->keras>=3.2.0-
>tensorflow<2.18,>=2.17->tf-keras) (2.19.1)
Requirement already satisfied: intel-cmplr-lib-ur==2024.2.0 in
/usr/local/lib/python3.10/dist-packages (from intel-openmp>=2024->mkl-
>numpy<2.0.0,>=1.23.5->tensorflow<2.18,>=2.17->tf-keras) (2024.2.0)
Requirement already satisfied: mdurl~=0.1 in
/usr/local/lib/python3.10/dist-packages (from markdown-it-py>=2.2.0-
>rich->keras>=3.2.0->tensorflow<2.18,>=2.17->tf-keras) (0.1.2)
```

Import Libs

```
%matplotlib inline
import pandas
import sklearn
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

from torch.utils.data import Dataset
from IPython.display import display
from collections import defaultdict
from sklearn.metrics import accuracy_score

data_df = pd.read_csv('clean-phone-data-for-students.csv')

def clean_data(df):
    """Cleans the dataset by selecting relevant columns, normalizing
labels,
    trimming whitespace, and removing duplicates."""

    # Select and rename columns
    df = df[["Sentence Utterance",
"Object"]].rename(columns={"Sentence Utterance": "input", "Object":
"raw_label"})

    # Normalize label (lowercase)
    df["clean_label"] = df["raw_label"].str.lower()
```

```
    # Trim white spaces in input column
    df["input"] = df["input"].str.strip()

    # Remove duplicates based on input
    df = df.drop_duplicates(subset="input", keep="first")

    # Drop the raw label column
    df.drop(columns=["raw_label"], inplace=True)

    return df

# Apply cleaning function
data_df = clean_data(data_df)

# Display summary
display(data_df.describe())
display(data_df["clean_label"].unique())
```

	input	clean_label
count	13367	13367
unique	13367	26
top	สอบถามโปรโมชั่นปัจจุบันที่ใช้อยู่ค่ะ	service
freq	1	2108

```
array(['payment', 'package', 'suspend', 'internet', 'phone_issues',
       'service', 'nontruemove', 'balance', 'detail', 'bill',
'credit',
       'promotion', 'mobile_setting', 'iservice', 'roaming',
'truemoney',
       'information', 'lost_stolen', 'balance_minutes', 'idd',
'garbage',
       'ringtone', 'rate', 'loyalty_card', 'contact', 'officer'],
      dtype=object)
```

```
# Mapping and Trimming
data = data_df.to_numpy()
unique_label = data_df.clean_label.unique()

label_2_num_map = dict(zip(unique_label, range(len(unique_label))))
num_2_label_map = dict(zip(range(len(unique_label)), unique_label))

data[:,1] = np.vectorize(label_2_num_map.get)(data[:,1])

def strip_str(string):
    return string.strip()
data[:,0] = np.vectorize(strip_str)(data[:,0])

display(data)
```

```
array([['<PHONE_NUMBER_REMOVED> ผมไปจ่ายเงินที่ Counter Services เค้าเช็ต
3276.25 บาท เมื่อวานที่ผมเช็คที่ศูนย์บอกมียอด 3057.79 บาท',
        0],
       ['internet ยังความเร็วอยู่เท่าไหร ครับ', 1],
       ['ตะกี้ไปชำระค่าบริการไปแล้ว แต่ยังใช้งานไม่ได้ ค่ะ', 2],
       ...,
       ['ยอดเงินเหลือเท่าไหร่ค่ะ', 7],
       ['ยอดเงินในระบบ', 7],
       ['สอบถามโปรโมชั่นปัจจุบันที่ใช้อยู่ค่ะ', 1]], dtype=object)

# Split
from sklearn.model_selection import train_test_split

# Constants
SEED = 42
MIN_INSTANCES = 10  # Minimum instances per class


def filter_data(data_df, min_instances=MIN_INSTANCES):
    """
    Filters classes with fewer than `min_instances` occurrences.
    Returns filtered input (X) and labels (y).
    """
    class_counts = data_df["clean_label"].value_counts()
    valid_classes = class_counts[class_counts >= min_instances].index

    filtered_data =
data_df[data_df["clean_label"].isin(valid_classes)]
    return filtered_data["input"],
filtered_data["clean_label"].astype(int)

def split_data(data_df, random_state=SEED,
min_instances=MIN_INSTANCES):
    """
    Splits data into train (80%), validation (10%), and test (10%)
sets.
    Ensures stratification and filtering of rare classes.
    """
    # Filter classes
    X, y = filter_data(data_df, min_instances)

    # Split 80% Train, 20% Temp
    X_train, X_temp, y_train, y_temp = train_test_split(
        X, y, test_size=0.20, stratify=y, random_state=random_state
    )

    # Split 10% Validation, 10% Test
    X_val, X_test, y_val, y_test = train_test_split(
        X_temp, y_temp, test_size=0.50, stratify=y_temp,
```

```
random_state=random_state
    )

    print(f"Train size: {len(X_train)}")
    print(f"Validation size: {len(X_val)}")
    print(f"Test size: {len(X_test)}")

    return (
        np.array(X_train), np.array(X_val), np.array(X_test),
        np.array(y_train), np.array(y_val), np.array(y_test)
    )

# Convert to DataFrame
df = pd.DataFrame(data, columns=['input', 'clean_label'])

# Split dataset
X_train, X_val, X_test, y_train, y_val, y_test = split_data(df)
```

```
Train size: 10690
Validation size: 1336
Test size: 1337
```

Model 2 MUSE

Build a simple logistic regression model using features from the MUSE model.

Which MUSE model will you use? Why?

Ans: I will use sentence-transformers/use-cmlm-multilingual because:

- It is pre-trained on multiple languages, including Thai, ensuring better language coverage.
- It captures sentence-level semantics rather than just individual words, leading to more meaningful embeddings.
- It generalizes better than traditional vector-based models like TF-IDF, improving performance in downstream tasks.

MUSE is typically used with tensorflow. However, there are some pytorch conversions made by some people.

https://huggingface.co/sentence-transformers/use-cmlm-multilingual
https://huggingface.co/dayyass/universal-sentence-encoder-multilingual-large-3-pytorch

```
from sentence_transformers import SentenceTransformer
from sklearn.linear_model import LogisticRegression
start_time = time.time()
print("MUSE + Logistic Regression")

muse_model = SentenceTransformer("sentence-transformers/use-cmlm-
```

```
multilingual")

def encode_texts(texts):
    return muse_model.encode(texts, convert_to_numpy=True)

start_enc_time = time.time()
X_train_enc = encode_texts(X_train.tolist())
X_val_enc = encode_texts(X_val.tolist())
X_test_enc = encode_texts(X_test.tolist())
end_enc_time = time.time()
print(f"Encoding Time: {end_enc_time - start_enc_time:.4f} seconds")

model = LogisticRegression(random_state=SEED)

start_train_time = time.time()
model.fit(X_train_enc, y_train)
end_train_time = time.time()
print(f"Training Time: {end_train_time - start_train_time:.4f}
seconds")

y_pred_train = model.predict(X_train_enc)
y_pred_val = model.predict(X_val_enc)
y_pred_test = model.predict(X_test_enc)

train_acc = np.mean(y_train.astype(int) == y_pred_train)
val_acc = np.mean(y_val.astype(int) == y_pred_val)
test_acc = np.mean(y_test.astype(int) == y_pred_test)

print(f"Train Accuracy: {train_acc:.4f}")
print(f"Validation Accuracy: {val_acc:.4f}")
print(f"Test Accuracy: {test_acc:.4f}")
end_time = time.time()
print(f"Total Time: {end_time - start_time:.4f} seconds")
```

MUSE + Logistic Regression


```
{"model_id":"989b24c6231f4c1897190e1e8bbf71df","version_major":2,"vers
ion_minor":0}

{"model_id":"48248e8130354129b17eb409cda3a458","version_major":2,"vers
ion_minor":0}

{"model_id":"d356c37e628c4511b4ef235f46ea564b","version_major":2,"vers
ion_minor":0}

{"model_id":"4e91200f367c4930861e909c3cbf30ad","version_major":2,"vers
ion_minor":0}

{"model_id":"eb15376b321d4203beb31dceec1e4de5","version_major":2,"vers
ion_minor":0}
```

{"model_id":"ed2161d3aad04da0b4fc86e0ca4616a1","version_major":2,"version_minor":0}

```
Some weights of the model checkpoint at sentence-transformers/use-cmlm-multilingual were not used when initializing BertModel: ['cls.predictions.bias', 'cls.predictions.transform.LayerNorm.bias', 'cls.predictions.transform.LayerNorm.weight', 'cls.predictions.transform.dense.bias', 'cls.predictions.transform.dense.weight', 'cls.seq_relationship.bias', 'cls.seq_relationship.weight']
- This IS expected if you are initializing BertModel from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing BertModel from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).
```

{"model_id":"e553f244937f479ba514ba4d4b532531","version_major":2,"version_minor":0}

{"model_id":"44cd7245ed624b5cbb978752ae579534","version_major":2,"version_minor":0}

{"model_id":"7a6ecf76c16d42be9d97b9a9be7b88cd","version_major":2,"version_minor":0}

{"model_id":"c2669f8819ad4bb7b29ef6cbc1e00adc","version_major":2,"version_minor":0}

{"model_id":"99eb8e02259841b6a67a2be0c1f288ce","version_major":2,"version_minor":0}

{"model_id":"4d0e527b97104cf7891ab99b88962718","version_major":2,"version_minor":0}

{"model_id":"37b7df74d03e4631b63dcee7d1e14789","version_major":2,"version_minor":0}

{"model_id":"975c55649cbc4ea9a727b5a7685de82a","version_major":2,"version_minor":0}

```
Encoding Time: 21.6718 seconds
Training Time: 2.2648 seconds
Train Accuracy: 0.7373
Validation Accuracy: 0.7073
Test Accuracy: 0.7023
Total Time: 38.5585 seconds

/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed to converge
```

```
(status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as
shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

Comparison

After you have completed the 3 models, compare the accuracy, ease of implementation, and inference speed (from cleaning, tokenization, till model compute) between the three models in mycourseville.