In [6]:
```
!pip install transformers datasets scikit-learn pandas matplotlib seaborn
```

```
Requirement already satisfied: transformers in c:\users\arthu\appdata\loca
l\programs\python\python311\lib\site-packages (4.35.0)
Requirement already satisfied: datasets in c:\users\arthu\appdata\local\pr
ograms\python\python311\lib\site-packages (2.14.7)
Requirement already satisfied: scikit-learn in c:\users\arthu\appdata\loca
l\programs\python\python311\lib\site-packages (1.3.2)
Requirement already satisfied: pandas in c:\users\arthu\appdata\local\prog
rams\python\python311\lib\site-packages (2.1.4)
Requirement already satisfied: matplotlib in c:\users\arthu\appdata\local
\programs\python\python311\lib\site-packages (3.10.0)
Requirement already satisfied: seaborn in c:\users\arthu\appdata\local\pro
grams\python\python311\lib\site-packages (0.13.2)
Requirement already satisfied: filelock in c:\users\arthu\appdata\local\pr
ograms\python\python311\lib\site-packages (from transformers) (3.19.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.16.4 in c:\users\ar
thu\appdata\local\programs\python\python311\lib\site-packages (from transf
ormers) (0.17.3)
Requirement already satisfied: numpy>=1.17 in c:\users\arthu\appdata\local
\programs\python\python311\lib\site-packages (from transformers) (1.26.2)
Requirement already satisfied: packaging>=20.0 in c:\users\arthu\appdata\l
ocal\programs\python\python311\lib\site-packages (from transformers) (25.
0)
Requirement already satisfied: pyyaml>=5.1 in c:\users\arthu\appdata\local
\programs\python\python311\lib\site-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from transformers) (20
25.7.34)
Requirement already satisfied: requests in c:\users\arthu\appdata\local\pr
ograms\python\python311\lib\site-packages (from transformers) (2.31.0)
Requirement already satisfied: tokenizers<0.15,>=0.14 in c:\users\arthu\ap
pdata\local\programs\python\python311\lib\site-packages (from transformer
s) (0.14.1)
Requirement already satisfied: safetensors>=0.3.1 in c:\users\arthu\appdat
a\local\programs\python\python311\lib\site-packages (from transformers)
(0.6.2)
Requirement already satisfied: tqdm>=4.27 in c:\users\arthu\appdata\local
\programs\python\python311\lib\site-packages (from transformers) (4.67.1)
Requirement already satisfied: pyarrow>=8.0.0 in c:\users\arthu\appdata\lo
cal\programs\python\python311\lib\site-packages (from datasets) (22.0.0)
Requirement already satisfied: pyarrow-hotfix in c:\users\arthu\appdata\lo
cal\programs\python\python311\lib\site-packages (from datasets) (0.7)
Requirement already satisfied: dill<0.3.8,>=0.3.0 in c:\users\arthu\appdat
a\local\programs\python\python311\lib\site-packages (from datasets) (0.3.
7)
Requirement already satisfied: xxhash in c:\users\arthu\appdata\local\prog
rams\python\python311\lib\site-packages (from datasets) (3.6.0)
Requirement already satisfied: multiprocess in c:\users\arthu\appdata\loca
l\programs\python\python311\lib\site-packages (from datasets) (0.70.15)
Requirement already satisfied: fsspec<=2023.10.0,>=2023.1.0 in c:\users\ar
thu\appdata\local\programs\python\python311\lib\site-packages (from fsspec
[http]<=2023.10.0,>=2023.1.0->datasets) (2023.10.0)
Requirement already satisfied: aiohttp in c:\users\arthu\appdata\local\pro
grams\python\python311\lib\site-packages (from datasets) (3.13.3)
Requirement already satisfied: scipy>=1.5.0 in c:\users\arthu\appdata\loca
l\programs\python\python311\lib\site-packages (from scikit-learn) (1.15.3)
Requirement already satisfied: joblib>=1.1.1 in c:\users\arthu\appdata\loc
al\programs\python\python311\lib\site-packages (from scikit-learn) (1.5.2)
```

```
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\arthu\appd
ata\local\programs\python\python311\lib\site-packages (from scikit-learn)
(3.6.0)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\arthu\ap
pdata\local\programs\python\python311\lib\site-packages (from pandas) (2.
9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\arthu\appdata\loca
l\programs\python\python311\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.1 in c:\users\arthu\appdata\lo
cal\programs\python\python311\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from matplotlib) (1.3.
3)
Requirement already satisfied: cycler>=0.10 in c:\users\arthu\appdata\loca
l\programs\python\python311\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from matplotlib) (4.6
0.1)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from matplotlib) (1.4.
9)
Requirement already satisfied: pillow>=8 in c:\users\arthu\appdata\local\p
rograms\python\python311\lib\site-packages (from matplotlib) (10.0.1)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from matplotlib) (3.2.
5)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in c:\users\arthu\a
ppdata\local\programs\python\python311\lib\site-packages (from aiohttp->da
tasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.4.0 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from aiohttp->dataset
s) (1.4.0)
Requirement already satisfied: attrs>=17.3.0 in c:\users\arthu\appdata\loc
al\programs\python\python311\lib\site-packages (from aiohttp->datasets) (2
5.4.0)
Requirement already satisfied: frozenlist>=1.1.1 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from aiohttp->dataset
s) (1.8.0)
Requirement already satisfied: multidict<7.0,>=4.5 in c:\users\arthu\appda
ta\local\programs\python\python311\lib\site-packages (from aiohttp->datase
ts) (6.7.1)
Requirement already satisfied: propcache>=0.2.0 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from aiohttp->dataset
s) (0.4.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in c:\users\arthu\appdata
\local\programs\python\python311\lib\site-packages (from aiohttp->dataset
s) (1.22.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\arth
u\appdata\local\programs\python\python311\lib\site-packages (from huggingf
ace-hub<1.0,>=0.16.4->transformers) (4.15.0)
Requirement already satisfied: six>=1.5 in c:\users\arthu\appdata\local\pr
ograms\python\python311\lib\site-packages (from python-dateutil>=2.8.2->pa
ndas) (1.17.0)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\arthu
\appdata\local\programs\python\python311\lib\site-packages (from requests-
>transformers) (3.4.3)
Requirement already satisfied: idna<4,>=2.5 in c:\users\arthu\appdata\loca
l\programs\python\python311\lib\site-packages (from requests->transformer
s) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\arthu\appdat
```

```
a\local\programs\python\python311\lib\site-packages (from requests->transf
ormers) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\arthu\appdat
a\local\programs\python\python311\lib\site-packages (from requests->transf
ormers) (2025.8.3)
Requirement already satisfied: colorama in c:\users\arthu\appdata\local\pr
ograms\python\python311\lib\site-packages (from tqdm>=4.27->transformers)
(0.4.6)
[notice] A new release of pip is available: 24.0 -> 26.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

In [7]:
```python
import pandas as pd
import numpy as np
import re
import ast
import torch
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, f1_s
from transformers import AutoTokenizer, AutoModel
import matplotlib.pyplot as plt
import seaborn as sns
```

In [8]:
```python
df = pd.read_csv('SA_SubTxt_fn.csv')

# Clean the data column (it contains strings of lists)
def extract_text(text):
    try:
        if text.startswith('['):
            return ast.literal_eval(text)[0]
        return text
    except:
        return text

df['clean_text'] = df['data'].apply(extract_text)

# Multi-class Label Mapping (Mapping Ham to Personal, Support, Promotions
# 1 is Spam. 0 is Ham. We split 0 based on keywords.
def map_multi_class(row):
    text = str(row['clean_text']).lower()
    if row['label'] == 1:
        return 'Spam'
    elif any(kw in text for kw in ['support', 'help', 'issue', 'ticket',
        return 'Support'
    elif any(kw in text for kw in ['sale', 'offer', 'discount', 'price',
        return 'Promotions'
    else:
        return 'Personal'

df['category'] = df.apply(map_multi_class, axis=1)

print("Label Distribution:")
print(df['category'].value_counts())
```

```
Label Distribution:
category
Personal      2635
Spam          1896
Support       1349
```

```
Promotions     166
Name: count, dtype: int64
```

In [9]:
```python
def preprocess_text(text):
    text = text.lower()
    text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILIN
    text = re.sub(r'\@\w+|\#','', text) # Remove mentions/hashtags
    text = re.sub(r'[^a-z\s]', '', text) # Remove special chars and numbe
    return text

df['processed_text'] = df['clean_text'].apply(preprocess_text)
X_train, X_test, y_train, y_test = train_test_split(df['processed_text'],
```

In [10]:
```python
tfidf = TfidfVectorizer(max_features=5000)
X_train_tfidf = tfidf.fit_transform(X_train)
X_test_tfidf = tfidf.transform(X_test)
```

In [11]:
```python
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train_tfidf, y_train)
y_pred_rf = rf_model.predict(X_test_tfidf)
```

In [12]:
```python
tokenizer = AutoTokenizer.from_pretrained("distilbert-base-uncased")
model = AutoModel.from_pretrained("distilbert-base-uncased")

def get_embeddings(text_list):
    inputs = tokenizer(text_list.tolist(), padding=True, truncation=True,
    with torch.no_grad():
        outputs = model(**inputs)
    # Use the mean of hidden states as sentence representation
    return outputs.last_hidden_state.mean(dim=1).numpy()

# Note: For large datasets, process in batches. Here we use a sample for
sample_size = 500
X_test_embeddings = get_embeddings(X_test[:sample_size])
print("GenAI Embeddings Shape:", X_test_embeddings.shape)
```

```
Downloading tokenizer_config.json:   0%|          | 0.00/48.0 [00:00<?, ?
B/s]
C:\Users\arthu\AppData\Local\Programs\Python\Python311\Lib\site-packages\h
uggingface_hub\file_download.py:137: UserWarning: `huggingface_hub` cache-
system uses symlinks by default to efficiently store duplicated files but
your machine does not support them in C:\Users\arthu\.cache\huggingface\hu
b. Caching files will still work but in a degraded version that might requ
ire more space on your disk. This warning can be disabled by setting the `
HF_HUB_DISABLE_SYMLINKS_WARNING` environment variable. For more details, s
ee https://huggingface.co/docs/huggingface_hub/how-to-cache#limitations.
To support symlinks on Windows, you either need to activate Developer Mode
or to run Python as an administrator. In order to see activate developer m
ode, see this article: https://docs.microsoft.com/en-us/windows/apps/get-s
tarted/enable-your-device-for-development
  warnings.warn(message)
Downloading config.json:   0%|          | 0.00/483 [00:00<?, ?B/s]
Downloading vocab.txt:   0%|          | 0.00/232k [00:00<?, ?B/s]
Downloading tokenizer.json:   0%|          | 0.00/466k [00:00<?, ?B/s]
Downloading model.safetensors:   0%|          | 0.00/268M [00:00<?, ?B/s]
GenAI Embeddings Shape: (500, 768)
```

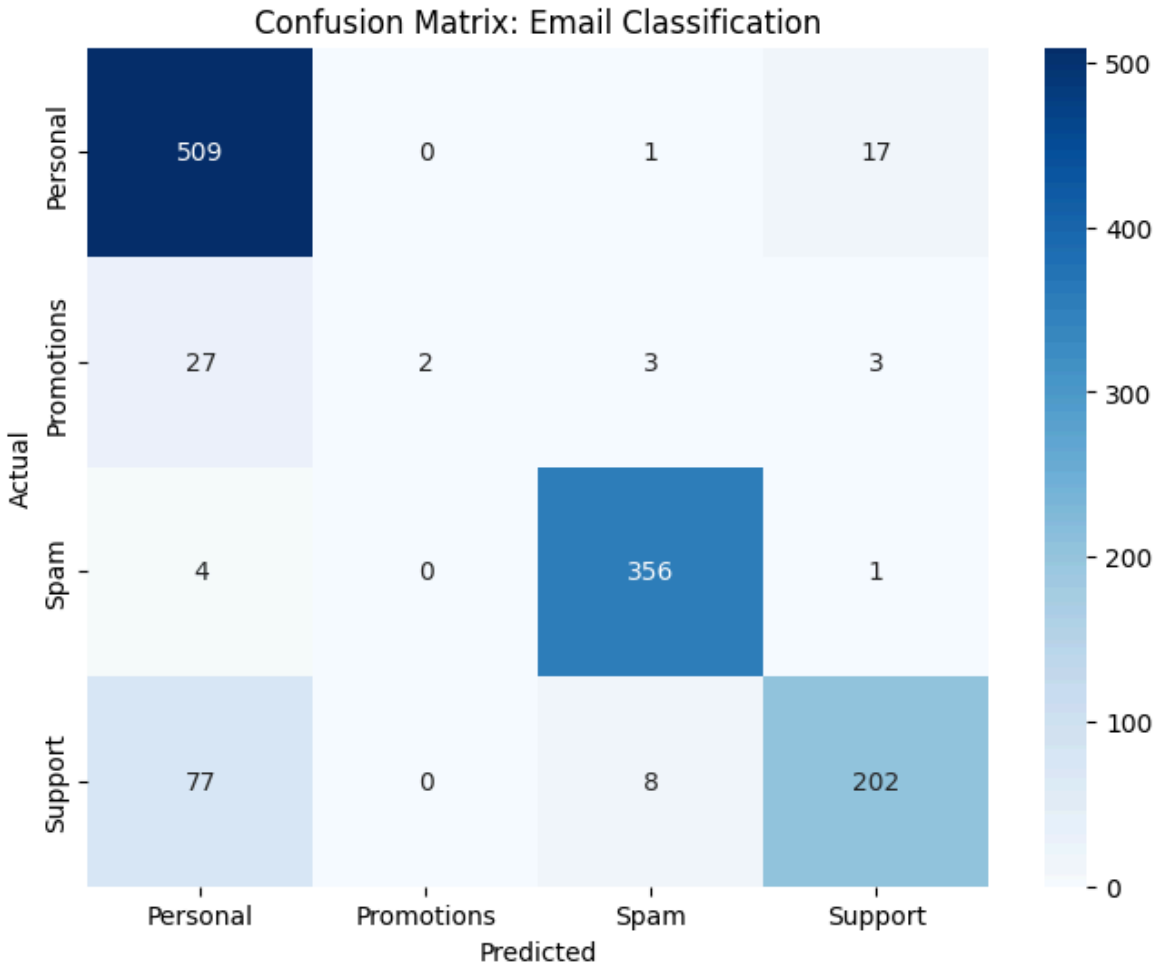
In [13]:
```python
print("\n--- Classification Report (TF-IDF + Random Forest) ---")
print(classification_report(y_test, y_pred_rf))
```

```
# Confusion Matrix Visualization
cm = confusion_matrix(y_test, y_pred_rf)
plt.figure(figsize=(8,6))
sns.heatmap(cm, annot=True, fmt='d', xticklabels=rf_model.classes_, ytick
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix: Email Classification')
plt.show()
```

```
--- Classification Report (TF-IDF + Random Forest) ---
              precision    recall  f1-score   support

    Personal       0.82      0.97      0.89       527
  Promotions       1.00      0.06      0.11        35
        Spam       0.97      0.99      0.98       361
     Support       0.91      0.70      0.79       287

    accuracy                           0.88      1210
   macro avg       0.92      0.68      0.69      1210
weighted avg       0.89      0.88      0.87      1210
```



In [14]:

```
def predict_email(email_text):
    proc = preprocess_text(email_text)
    vec = tfidf.transform([proc])
    return rf_model.predict(vec)[0]

test_email = "Get 50% off on your next purchase! Limited time offer."
print(f"Sample Email: {test_email}")
print(f"Predicted Category: {predict_email(test_email)}")
```

Sample Email: Get 50% off on your next purchase! Limited time offer.
Predicted Category: Spam