2. Develop a text classification model that can effectively identify, extract features, and classify documents from the 20 Newsgroups dataset into one of the 20 predefined categories using pattern recognition techniques.

```
In [1]: import pandas as pd
         from sklearn.datasets import fetch_20newsgroups
         from sklearn.model_selection import train_test_split
         from sklearn.preprocessing import MaxAbsScaler
         from sklearn.feature extraction.text import TfidfVectorizer
         from sklearn.linear_model import LogisticRegression
         from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
         import nltk
         from nltk.corpus import stopwords
         import re
         import matplotlib.pyplot as plt
         import seaborn as sns
 In [3]: newsgroups = fetch_20newsgroups(subset='all', shuffle=True, random_state=42)
 In [4]: nltk.download('stopwords')
         stop_words = stopwords.words('english')
        [nltk_data] Downloading package stopwords to
        [nltk_data]
                        C:\Users\ASUS\AppData\Roaming\nltk_data...
        [nltk_data] Unzipping corpora\stopwords.zip.
 In [7]: def preprocess_text(text):
             text = text.lower()
             text = re.sub(r'[^a-zA-Z\s]', '', text)
             stop_words = set(stopwords.words('english'))
             tokens = text.split()
             tokens = [word for word in tokens if word not in stop_words]
             return ' '.join(tokens)
In [12]: | newsgroups_data = pd.DataFrame({'text': newsgroups.data, 'target': newsgroups.target})
         newsgroups_data['text'] = newsgroups_data['text'].apply(preprocess_text)
In [13]: newsgroups_data.head()
Out[13]:
                                                       text target
            mamatha devineni ratnam mrandrewcmuedu subject...
                                                               10
          1 mblawsonmidwayecnuoknoredu matthew b lawson su...
                                                                3
          2
                   hilmierdsvsuse hilmi eren subject armenia says...
                                                               17
         3
                guydaustinibmcom guy dawson subject ide vs scs...
             alexander samuel mcdiarmid amoandrewcmuedu sub...
                                                                4
In [14]: X_train, X_test, y_train, y_test = train_test_split(
             newsgroups_data['text'],
             newsgroups_data['target'],
             test_size=0.2,
             random_state=42
In [15]: vectorizer = TfidfVectorizer(stop_words=stop_words, max_features=5000)
         X_train_vec = vectorizer.fit_transform(X_train)
         X_test_vec = vectorizer.transform(X_test)
In [16]: scaler = MaxAbsScaler()
         X_train = scaler.fit_transform(X_train_vec)
         X_test = scaler.transform(X_test_vec)
In [17]: model = LogisticRegression(max_iter=1000)
         model.fit(X_train, y_train)
Out[17]:
                LogisticRegression
         LogisticRegression(max_iter=1000)
In [24]: y_pred = model.predict(X_test)
         print(f"Accuracy: {accuracy_score(y_test, y_pred)}")
         print(classification_report(y_test, y_pred, target_names=newsgroups.target_names))
```

```
comp.graphics
                                               0.70
                                                           0.76
                                                                       0.73
                                                                                    202
           comp.os.ms-windows.misc
                                                                                    195
                                               0.81
                                                           0.81
                                                                       0.81
         comp.sys.ibm.pc.hardware
                                               0.66
                                                           0.71
                                                                       0.69
                                                                                    183
                                                           0.82
                                               0.84
                                                                       0.83
                                                                                    205
             comp.sys.mac.hardware
                     comp.windows.x
                                               0.81
                                                           0.81
                                                                       0.81
                                                                                    215
                        misc.forsale
                                                           0.79
                                               0.86
                                                                       0.82
                                                                                    193
                                                           0.91
                                                                       0.90
                                                                                    196
                            rec.autos
                                               0.89
                    rec.motorcycles
                                               0.97
                                                           0.92
                                                                       0.95
                                                                                    168
                rec.sport.baseball
                                               0.92
                                                           0.93
                                                                       0.92
                                                                                    211
                                               0.95
                                                           0.97
                                                                       0.96
                                                                                    198
                   rec.sport.hockey
                                               0.96
                                                           0.93
                                                                       0.94
                                                                                    201
                            sci.crypt
                                                           0.79
                                                                                    202
                    sci.electronics
                                               0.80
                                                                       0.80
                              sci.med
                                               0.91
                                                           0.90
                                                                       0.91
                                                                                    194
                                                           0.96
                                                                                    189
                            sci.space
                                               0.95
                                                                       0.96
            soc.religion.christian
                                                           0.96
                                                                       0.93
                                                                                    202
                                               0.90
                                               0.92
                 talk.politics.guns
                                                           0.91
                                                                       0.92
                                                                                    188
             talk.politics.mideast
                                               0.94
                                                           0.95
                                                                       0.94
                                                                                    182
                 talk.politics.misc
                                               0.86
                                                           0.87
                                                                       0.86
                                                                                    159
                                                           0.69
                                                                       0.76
                                                                                    136
                 talk.religion.misc
                                               0.85
                            accuracy
                                                                       0.86
                                                                                   3770
                                                                                   3770
                            macro avg
                                               0.87
                                                           0.86
                                                                       0.86
                        weighted avg
                                               0.87
                                                           0.86
                                                                       0.87
                                                                                   3770
In [19]:
           plt.figure(figsize=(10, 7))
           sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='Blues', xticklabels=newsgroups.target_names, yticklabels=newsgroup
           plt.xlabel('Predicted')
           plt.ylabel('Actual')
           plt.title('Confusion Matrix')
           plt.show()
                                                                                       Confusion Matrix
                                alt.atheism - 129
                                                           0
                                                                0
                                                                     0
                                                                          0
                                                                                0
                                                                                               1
                                                                                                     1
                                                                                                          0
                                                                                                               0
                                                                                                                    1
                                                                                                                                    0
                                                                                                                                         1
                                                                                                                                              2
                                                                                                                                                   12
                            comp.graphics - 0
                                                    153
                                                          14
                                                                                                                                         2
                                                                     1
                                                                          15
                                                                                0
                                                                                               3
                                                                                                     0
                                                                                                                         2
                                                                                                                                    0
                                                                                                                                              0
                                                                                                                                                    0
                                                                                                                                                                  - 175
                                                         158 16
                                                                     2
              comp.os.ms-windows.misc - 0
                                                     10
                                                                          8
                                                                                                                          0
                                                                                                                               0
                                                                                                                                    0
                                                                                                                                         0
                                                                                1
                                                                                               0
                                                                                                     0
                                                                                                          0
                                                                                                               0
                                                                                                                    0
                                                                                                                                              0
                                                                                                                                                    0
                                                                                     0
                                                                                          0
                                                                          7
             comp.sys.ibm.pc.hardware - 0
                                                     9
                                                           8
                                                               130
                                                                    12
                                                                                5
                                                                                          0
                                                                                               0
                                                                                                     0
                                                                                                          1
                                                                                                               9
                                                                                                                    1
                                                                                                                          0
                                                                                                                               0
                                                                                                                                    0
                                                                                                                                         0
                                                                                                                                              0
                                                                                                                                                    0
                                                                                     1
                comp.sys.mac.hardware - 0
                                                     3
                                                          3
                                                               11
                                                                    169
                                                                          3
                                                                                     0
                                                                                          1
                                                                                               1
                                                                                                     0
                                                                                                          0
                                                                                                               5
                                                                                                                    3
                                                                                                                          1
                                                                                                                              0
                                                                                                                                    0
                                                                                                                                         0
                                                                                                                                                    0
                                                                                                                                              1
                                                                                                                                                                  - 150
                                                                         174
                          comp.windows.x - 0
                                                     25
                                                           6
                                                                4
                                                                     1
                                                                                          0
                                                                                               2
                                                                                                                                    0
                                                                                                                                         1
                                                                                                                                              0
                                                                                                                                                    0
                                                                     5
                                                                               153
                                                                                               2
                               misc.forsale - 0
                                                     1
                                                           1
                                                               16
                                                                          0
                                                                                     4
                                                                                          1
                                                                                                     1
                                                                                                               6
                                                                                                                    1
                                                                                                                          1
                                                                                                                               0
                                                                                                                                    0
                                                                                                                                         0
                                                                                                                                              0
                                                                                                                                                    0
                                                                                                          1
                                                                                                                                                                  - 125
                                                                1
                                                                                3
                                                                                   178
                                                                                          1
                                                                                               2
                                                                                                     0
                                  rec.autos - 0
                                                     5
                                                           1
                                                                     0
                                                                          0
                                                                                                          0
                                                                                                               3
                                                                                                                    1
                                                                                                                          1
                                                                                                                              0
                                                                                                                                    0
                                                                                                                                         0
                                                                                                                                              0
                                                                                                                                                    0
                                                                                         155
                                                                                               0
                          rec.motorcycles - 0
                                                           0
                                                                1
                                                                     0
                                                                          1
                                                                                     3
                                                                                                     0
                                                                                                          0
                                                                                                               0
                                                                                                                               0
                                                                                                                                    1
                                                                                                                                         0
                                                                                                                                              1
                                                                                                                                                    0
                                                                                              197
                                                                                                    8
                                                                                                          0
                                                                                                               2
                        rec.sport.baseball - 2
                                                      0
                                                           0
                                                                0
                                                                     0
                                                                          0
                                                                                0
                                                                                     1
                                                                                                                    0
                                                                                                                          0
                                                                                                                              1
                                                                                                                                    0
                                                                                                                                         0
                                                                                                                                              0
                                                                                                                                                    0
          Actual
                                                                                                                                                                  - 100
                                                                                               1
                                                                                                   193
                                                                                                          0
                                                                                                               0
                         rec.sport.hockey - 0
                                                     0
                                                           0
                                                                0
                                                                     1
                                                                                1
                                                                                     1
                                                                                          0
                                                                                                                    0
                                                                                                                         0
                                                                                                                              1
                                                                                                                                    0
                                                                                                                                         0
                                                                                                                                              0
                                                                          0
                                                                                                                                                    0
                                   sci.crypt - 0
                                                      4
                                                           1
                                                                1
                                                                     0
                                                                          0
                                                                                1
                                                                                     0
                                                                                          0
                                                                                               0
                                                                                                     0
                                                                                                        186
                                                                                                              3
                                                                                                                    0
                                                                                                                         1
                                                                                                                              1
                                                                                                                                    0
                                                                                                                                         2
                                                                                                                                              1
                                                                                                                                                    0
                                                                                                                    2
                            sci.electronics - 0
                                                     3
                                                           1
                                                                8
                                                                     7
                                                                          3
                                                                                6
                                                                                     5
                                                                                          0
                                                                                               3
                                                                                                     0
                                                                                                          1
                                                                                                             160
                                                                                                                          1
                                                                                                                               0
                                                                                                                                    1
                                                                                                                                         1
                                                                                                                                              0
                                                                                                                                                    0
                                                                                                                                                                   - 75
                                                                                                          0
                                                                                                                   175
                                                                                                                          0
                                                                                                                               0
                                                                                                                                    0
                                    sci.med - 0
                                                     4
                                                           1
                                                                1
                                                                     1
                                                                          0
                                                                                0
                                                                                     4
                                                                                          0
                                                                                               0
                                                                                                     0
                                                                                                               6
                                                                                                                                         0
                                                                                                                                              1
                                                                                                                                                   1
                                                                                                                    2
                                                                                                                        181
                                                                                                                                    0
                                                                                                                                         0
                                  sci.space - 1
                                                     1
                                                           0
                                                                0
                                                                     1
                                                                          1
                                                                                               0
                                                                                                     0
                                                                                                          0
                                                                                                               0
                                                                                                                              1
                                                                                                                                              1
                                                                                                                                                    0
                                                                                0
                                                                                     0
                                                                                          0
                                                                                                                                                                   - 50
                                                                                                                             194
                                                                                                                                    0
                                                                                                                                         1
                                                                                                                                              0
                     soc.religion.christian - 2
                                                     0
                                                           0
                                                                0
                                                                     0
                                                                          1
                                                                                0
                                                                                          0
                                                                                               0
                                                                                                     0
                                                                                                          0
                                                                                                               0
                                                                                                                    0
                                                                                                                          0
                                                                                                                                                   3
                                                                                     1
                                                                                                                                   172
                         talk.politics.guns - 1
                                                     0
                                                           0
                                                                0
                                                                     0
                                                                          0
                                                                                0
                                                                                               1
                                                                                                     0
                                                                                                          1
                                                                                                               0
                                                                                                                    0
                                                                                                                              2
                                                                                                                                         1
                                                                                                                                              9
                                                                                                                                                   1
                                                                                     0
                                                                                          0
                                                                                                                          0
                     talk.politics.mideast - 2
                                                           0
                                                                          2
                                                                                                               1
                                                                                                                    0
                                                                                                                                                    0
                                                                0
                                                                     0
                                                                                0
                                                                                     0
                                                                                          0
                                                                                               0
                                                                                                     0
                                                                                                                                                                  - 25
                         talk.politics.misc - 1
                                                                0
                                                                     0
                                                                                                                    2
                                                      0
                                                           0
                                                                          0
                                                                                0
                                                                                     1
                                                                                          0
                                                                                               2
                                                                                                     1
                                                                                                          2
                                                                                                               0
                                                                                                                          1
                                                                                                                              2
                                                                                                                                         2
                                                                                                                                             138
                                                                                                                                                   0
                         talk.religion.misc - 18
                                                     0
                                                           0
                                                                0
                                                                     0
                                                                          0
                                                                                0
                                                                                          2
                                                                                               0
                                                                                                     0
                                                                                                          0
                                                                                                               1
                                                                                                                    0
                                                                                                                          0
                                                                                                                              11
                                                                                                                                    3
                                                                                                                                         0
                                                                                                                                                   94
                                                                                     0
                                                alt.atheism
                                                     comp.graphics
                                                                                                              sci.electronics
                                                                                                                   sci.med
                                                                                                                                             talk.politics.misc
                                                               comp.sys.ibm.pc.hardware
                                                                     comp.sys.mac.hardware
                                                                          comp.windows.x
                                                                                                         sci.crypt
                                                                                                                         sci.space
                                                                                                                              soc.religion.christian
                                                                                                                                   talk.politics.guns
                                                                                                                                        talk.politics.mideast
                                                                                                                                                   talk.religion.misc
                                                          comp.os.ms-windows.misc
                                                                                         rec.motorcycles
                                                                               misc.forsale
                                                                                    rec.autos
                                                                                              rec.sport.baseball
                                                                                                    rec.sport.hockey
                                                                                             Predicted
```

Accuracy: 0.8649867374005306

alt.atheism

precision

0.83

recall f1-score

0.84

0.85

support

In []: def predict_new_text(model, vectorizer, text):
 text_processed = preprocess_text(text)
 text_vectorized = vectorizer.transform([text_processed])
 prediction = model.predict(text_vectorized)
 # print(prediction)
 category_name = newsgroups.target_names[prediction[0]]

return category_name

In [22]: predicted_category = predict_new_text(model, vectorizer, "As the spacecraft drifted beyond Mars, its instruments captured breathtaking image print("As the spacecraft drifted beyond Mars, its instruments captured breathtaking images of the asteroid belt—countless ancient rocks float print(predicted_category)

[14]

As the spacecraft drifted beyond Mars, its instruments captured breathtaking images of the asteroid belt—countless ancient rocks floating in the silent vacuum, remnants of a solar system still shaping its story.
sci.space

In []: