nlp

November 5, 2024

1 Importing libraries

[20]:

```
import pandas as pd
import numpy as np
import re
import string
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, classification_report
import matplotlib.pyplot as plt
import seaborn as sns
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
import nltk
nltk.download('omw-1.4')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\criss\AppData\Roaming\nltk_data…
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\criss\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[20]:

```
True
```

2 Download nltk data (run once)

[14]:

```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\criss\AppData\Roaming\nltk_data…
[nltk_data]   Unzipping tokenizers\punkt.zip.
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\criss\AppData\Roaming\nltk_data…
[nltk_data]   Unzipping corpora\stopwords.zip.
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\criss\AppData\Roaming\nltk_data…
```

[14]:
```
True
```

[15]:
```
data = pd.read_csv(r"c:\flipitnews.csv")
```

[16]:
```
data
```

[16]:
```
           Category                                            Article
0        Technology  tv future in the hands of viewers with home th…
1          Business  worldcom boss  left books alone  former worldc…
2            Sports  tigers wary of farrell  gamble  leicester say …
3            Sports  yeading face newcastle in fa cup premiership s…
4     Entertainment  ocean s twelve raids box office ocean s twelve…
…             …                                                …
2220       Business  cars pull down us retail figures us retail sal…
2221       Politics  kilroy unveils immigration policy ex-chatshow …
2222  Entertainment  rem announce new glasgow concert us band rem h…
2223       Politics  how political squabbles snowball it s become c…
2224         Sports  souness delight at euro progress boss graeme s…

[2225 rows x 2 columns]
```

[17]:
```
print("Shape of the dataset:", data.shape)
print("Number of articles per category:")
print(data['Category'].value_counts())
```

```
Shape of the dataset: (2225, 2)
Number of articles per category:
Sports           511
Business         510
Politics         417
Technology       401
Entertainment    386
Name: Category, dtype: int64
```

3 Text Processing Function

[18]:
```python
def process_text(text):
    # Remove non-letter characters
    text = re.sub("[^a-zA-Z]", " ", text)
    # Tokenize text
    words = word_tokenize(text.lower())
    # Remove stopwords
    words = [word for word in words if word not in stopwords.words("english")]
    # Lemmatization
    lemmatizer = WordNetLemmatizer()
    words = [lemmatizer.lemmatize(word) for word in words]
    # Join words back into a single string
    return " ".join(words)
```

4 Apply text processing

[21]:
```python
data['Processed_Article'] = data['Article'].apply(process_text)
print("\nExample of processed article:")
print("Before:", data['Article'][0])
print("After:", data['Processed_Article'][0])
```

```

Example of processed article:
Before: tv future in the hands of viewers with home theatre systems  plasma
high-definition tvs  and digital video recorders moving into the living room
the way people watch tv will be radically different in five years  time.  that
is according to an expert panel which gathered at the annual consumer
electronics show in las vegas to discuss how these new technologies will impact
one of our favourite pastimes. with the us leading the trend  programmes and
other content will be delivered to viewers via home networks  through cable
satellite  telecoms companies  and broadband service providers to front rooms
and portable devices.  one of the most talked-about technologies of ces has been
digital and personal video recorders (dvr and pvr). these set-top boxes  like
the us s tivo and the uk s sky+ system  allow people to record  store  play
pause and forward wind tv programmes when they want.  essentially  the
technology allows for much more personalised tv. they are also being built-in to
high-definition tv sets  which are big business in japan and the us  but slower
to take off in europe because of the lack of high-definition programming. not
only can people forward wind through adverts  they can also forget about abiding
by network and channel schedules  putting together their own a-la-carte
entertainment. but some us networks and cable and satellite companies are
worried about what it means for them in terms of advertising revenues as well as
brand identity  and viewer loyalty to channels. although the us leads in this
technology at the moment  it is also a concern that is being raised in europe
particularly with the growing uptake of services like sky+.  what happens here
```

today we will see in nine months to a years time in the uk   adam hume the bbc broadcast s futurologist told the bbc news website. for the likes of the bbc there are no issues of lost advertising revenue yet. it is a more pressing issue at the moment for commercial uk broadcasters but brand loyalty is important for everyone. we will be talking more about content brands rather than network brands   said tim hanlon from brand communications firm starcom mediavest. the reality is that with broadband connections anybody can be the producer of content. he added: the challenge now is that it is hard to promote a programme with so much choice.   what this means said stacey jolna senior vice president of tv guide tv group is that the way people find the content they want to watch has to be simplified for tv viewers. it means that networks in us terms or channels could take a leaf out of google s book and be the search engine of the future instead of the scheduler to help people find what they want to watch. this kind of channel model might work for the younger ipod generation which is used to taking control of their gadgets and what they play on them. but it might not suit everyone the panel recognised. older generations are more comfortable with familiar schedules and channel brands because they know what they are getting. they perhaps do not want so much of the choice put into their hands mr hanlon suggested. on the other end you have the kids just out of diapers who are pushing buttons already - everything is possible and available to them   said mr hanlon. ultimately the consumer will tell the market they want.   of the 50 000 new gadgets and technologies being showcased at ces many of them are about enhancing the tv-watching experience. high-definition tv sets are everywhere and many new models of lcd (liquid crystal display) tvs have been launched with dvr capability built into them instead of being external boxes. one such example launched at the show is humax s 26-inch lcd tv with an 80-hour tivo dvr and dvd recorder. one of the us s biggest satellite tv companies directtv has even launched its own branded dvr at the show with 100-hours of recording capability instant replay and a search function. the set can pause and rewind tv for up to 90 hours. and microsoft chief bill gates announced in his pre-show keynote speech a partnership with tivo called tivotogo which means people can play recorded programmes on windows pcs and mobile devices. all these reflect the increasing trend of freeing up multimedia so that people can watch what they want when they want.
After: tv future hand viewer home theatre system plasma high definition tv digital video recorder moving living room way people watch tv radically different five year time according expert panel gathered annual consumer electronics show la vega discus new technology impact one favourite pastime u leading trend programme content delivered viewer via home network cable satellite telecom company broadband service provider front room portable device one talked technology ce digital personal video recorder dvr pvr set top box like u tivo uk sky system allow people record store play pause forward wind tv programme want essentially technology allows much personalised tv also built high definition tv set big business japan u slower take europe lack high definition programming people forward wind advert also forget abiding network channel schedule putting together la carte entertainment u network cable satellite company worried mean term advertising revenue well brand identity viewer loyalty channel although u lead technology moment also concern raised

```
europe particularly growing uptake service like sky happens today see nine month
year time uk adam hume bbc broadcast futurologist told bbc news website like bbc
issue lost advertising revenue yet pressing issue moment commercial uk
broadcaster brand loyalty important everyone talking content brand rather
network brand said tim hanlon brand communication firm starcom mediavest reality
broadband connection anybody producer content added challenge hard promote
programme much choice mean said stacey jolna senior vice president tv guide tv
group way people find content want watch simplified tv viewer mean network u
term channel could take leaf google book search engine future instead scheduler
help people find want watch kind channel model might work younger ipod
generation used taking control gadget play might suit everyone panel recognised
older generation comfortable familiar schedule channel brand know getting
perhaps want much choice put hand mr hanlon suggested end kid diaper pushing
button already everything possible available said mr hanlon ultimately consumer
tell market want new gadget technology showcased ce many enhancing tv watching
experience high definition tv set everywhere many new model lcd liquid crystal
display tv launched dvr capability built instead external box one example
launched show humax inch lcd tv hour tivo dvr dvd recorder one u biggest
satellite tv company directtv even launched branded dvr show hour recording
capability instant replay search function set pause rewind tv hour microsoft
chief bill gate announced pre show keynote speech partnership tivo called
tivotogo mean people play recorded programme window pc mobile device reflect
increasing trend freeing multimedia people watch want want
```

5 Encoding target variable

[24]:

```python
label_encoder = LabelEncoder()
data['Category_Label'] = label_encoder.fit_transform(data['Category'])
```

6 Feature extraction

[25]:

```python
def vectorize_data(method='tfidf'):
    if method == 'bow':
        vectorizer = CountVectorizer(max_features=5000)
    elif method == 'tfidf':
        vectorizer = TfidfVectorizer(max_features=5000)
    else:
        raise ValueError("Method should be 'bow' or 'tfidf'")
    X = vectorizer.fit_transform(data['Processed_Article']).toarray()
    y = data['Category_Label']
    return X, y
```

7 Choose vectorization method

[26]:

```
X, y = vectorize_data(method='tfidf')  # Change 'tfidf' to 'bow' for Bag of␣
 ↪Words
```

8 Train-test split

[27]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,␣
 ↪random_state=42)
print("Train shape:", X_train.shape)
print("Test shape:", X_test.shape)
```

```
Train shape: (1668, 5000)
Test shape: (557, 5000)
```

9 Train and Evaluate Models

[28]:

```
def evaluate_model(model, model_name):
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    print(f"\n{model_name} Classification Report:")
    print(classification_report(y_test, y_pred, target_names=label_encoder.
 ↪classes_))
    cm = confusion_matrix(y_test, y_pred)
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',␣
 ↪xticklabels=label_encoder.classes_, yticklabels=label_encoder.classes_)
    plt.title(f"{model_name} Confusion Matrix")
    plt.xlabel("Predicted")
    plt.ylabel("Actual")
    plt.show()
```

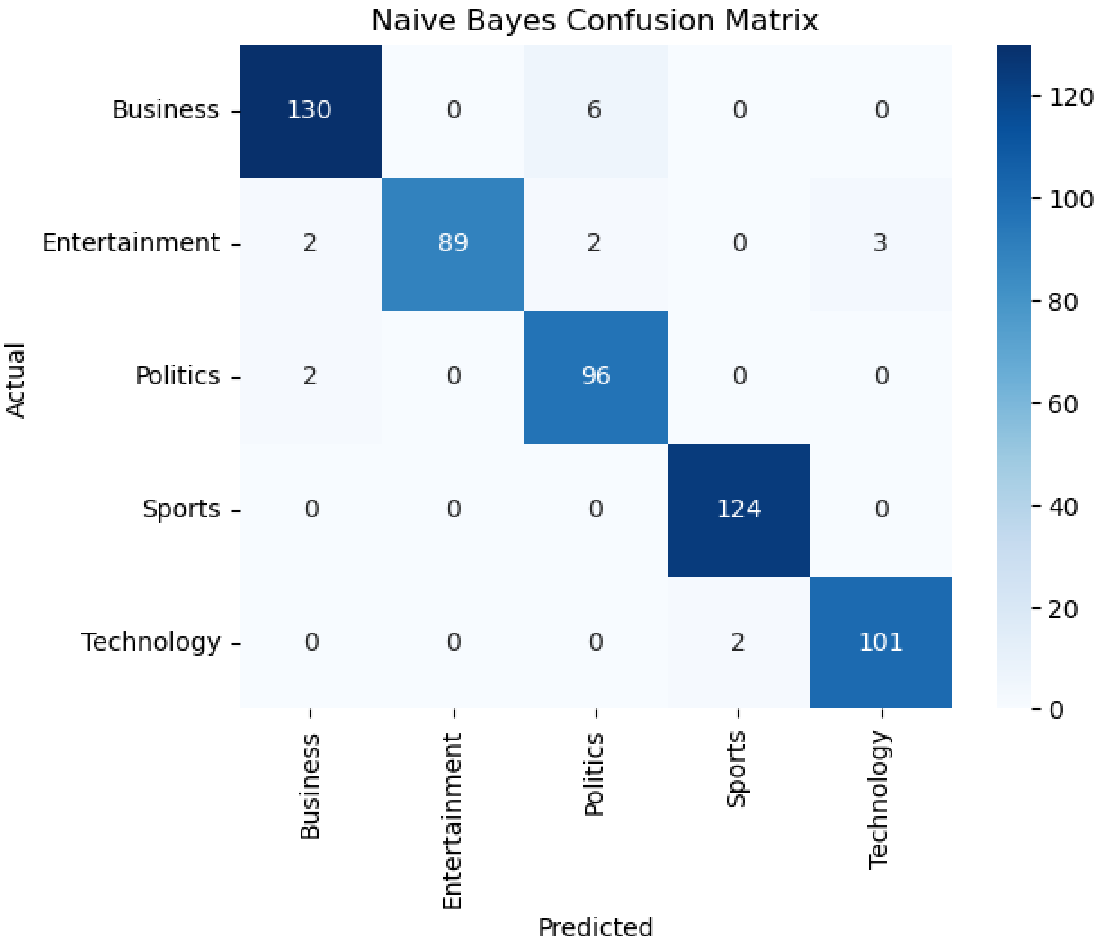
10 Naive Bayes Model

[29]:

```
nb_model = MultinomialNB()
evaluate_model(nb_model, "Naive Bayes")
```

```

Naive Bayes Classification Report:
               precision    recall  f1-score   support

     Business       0.97      0.96      0.96       136
Entertainment       1.00      0.93      0.96        96
     Politics       0.92      0.98      0.95        98
       Sports       0.98      1.00      0.99       124
   Technology       0.97      0.98      0.98       103
```

```
     accuracy                           0.97       557
    macro avg       0.97      0.97      0.97       557
 weighted avg       0.97      0.97      0.97       557
```

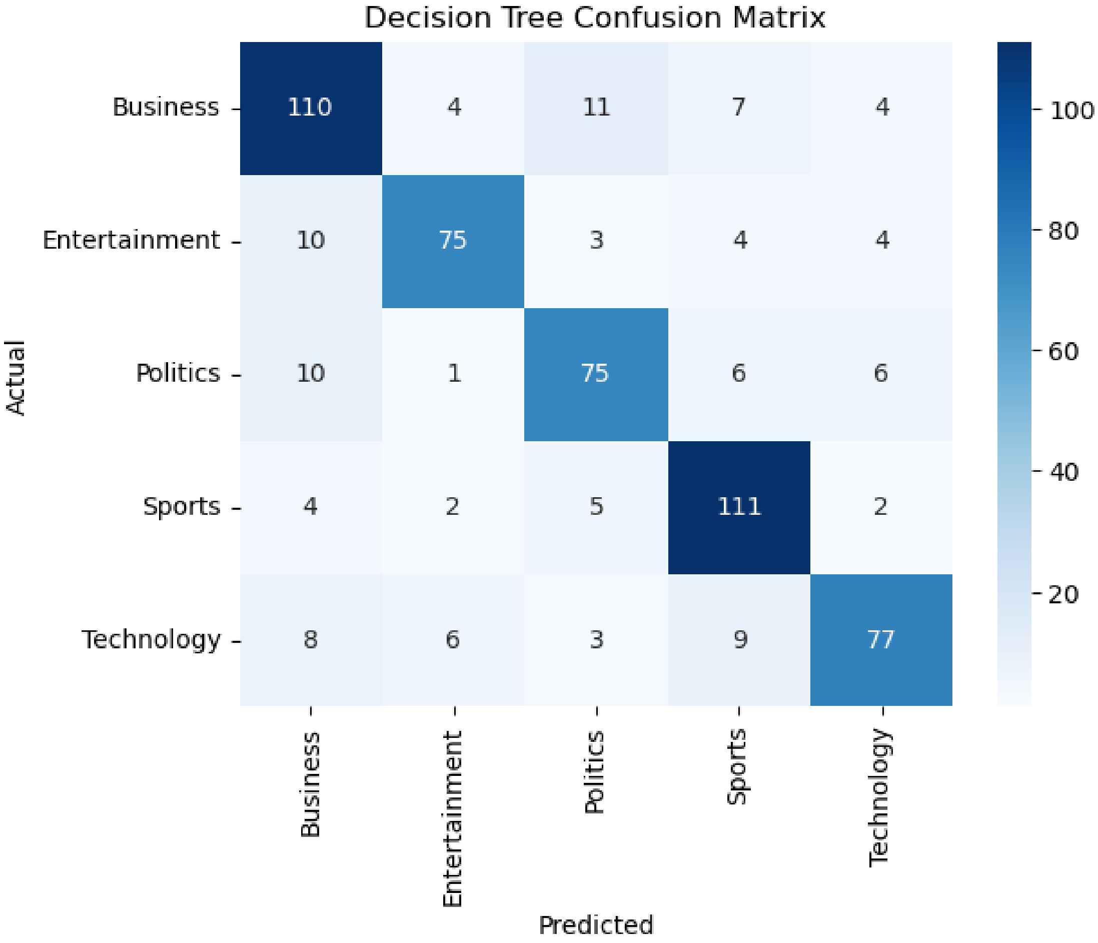


11 Decision Tree Model

[30]:
```
dt_model = DecisionTreeClassifier(random_state=42)
evaluate_model(dt_model, "Decision Tree")
```

```
Decision Tree Classification Report:
               precision    recall  f1-score   support

     Business       0.77      0.81      0.79       136
Entertainment       0.85      0.78      0.82        96
```

```
     Politics       0.77      0.77      0.77        98
       Sports       0.81      0.90      0.85       124
   Technology       0.83      0.75      0.79       103

     accuracy                           0.80       557
    macro avg       0.81      0.80      0.80       557
 weighted avg       0.81      0.80      0.80       557
```

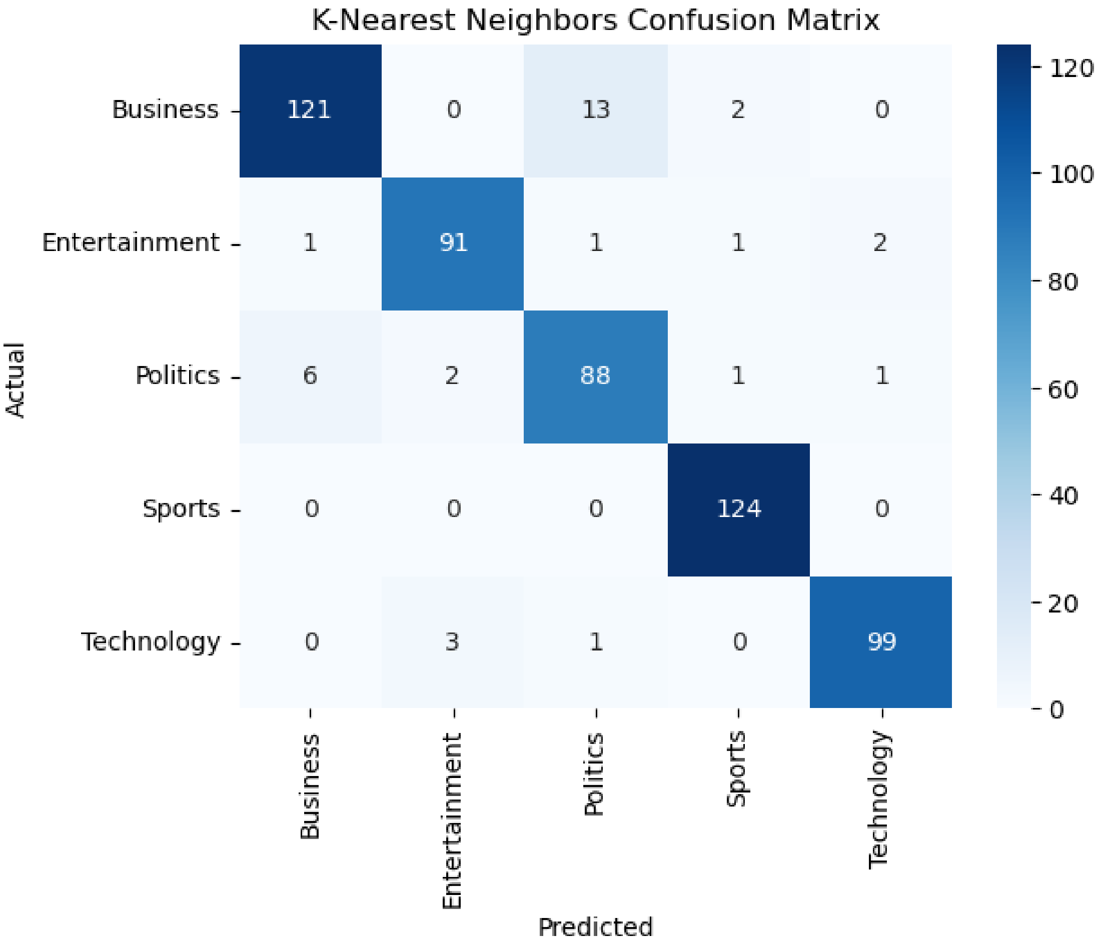


12 K-Nearest Neighbors Model

[31]:
```
knn_model = KNeighborsClassifier(n_neighbors=5)
evaluate_model(knn_model, "K-Nearest Neighbors")
```

```
K-Nearest Neighbors Classification Report:
               precision    recall  f1-score   support
```

```
     Business       0.95      0.89      0.92       136
Entertainment       0.95      0.95      0.95        96
     Politics       0.85      0.90      0.88        98
       Sports       0.97      1.00      0.98       124
   Technology       0.97      0.96      0.97       103

     accuracy                           0.94       557
    macro avg       0.94      0.94      0.94       557
 weighted avg       0.94      0.94      0.94       557
```


K-Nearest Neighbors Confusion Matrix

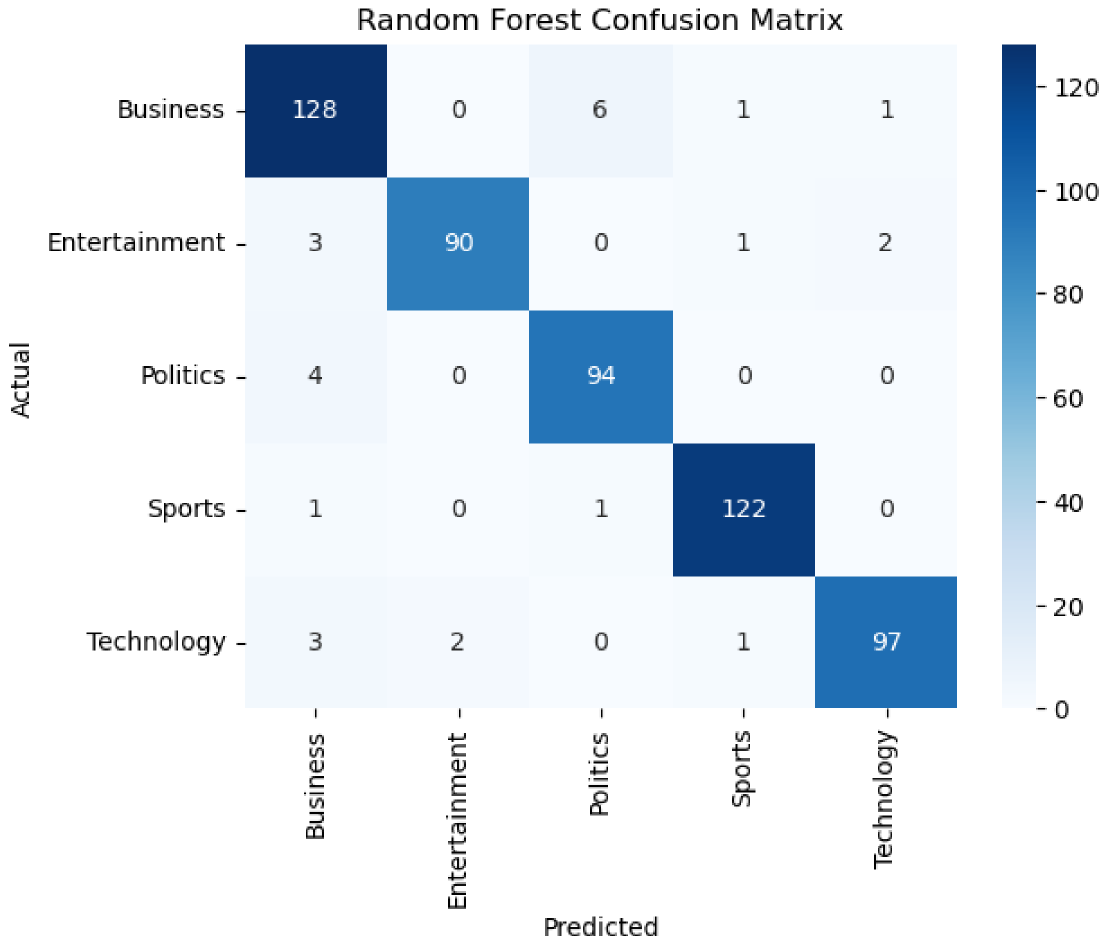
	Business	Entertainment	Politics	Sports	Technology
Business	121	0	13	2	0
Entertainment	1	91	1	1	2
Politics	6	2	88	1	1
Sports	0	0	0	124	0
Technology	0	3	1	0	99

13 Random Forest Model

[32]:

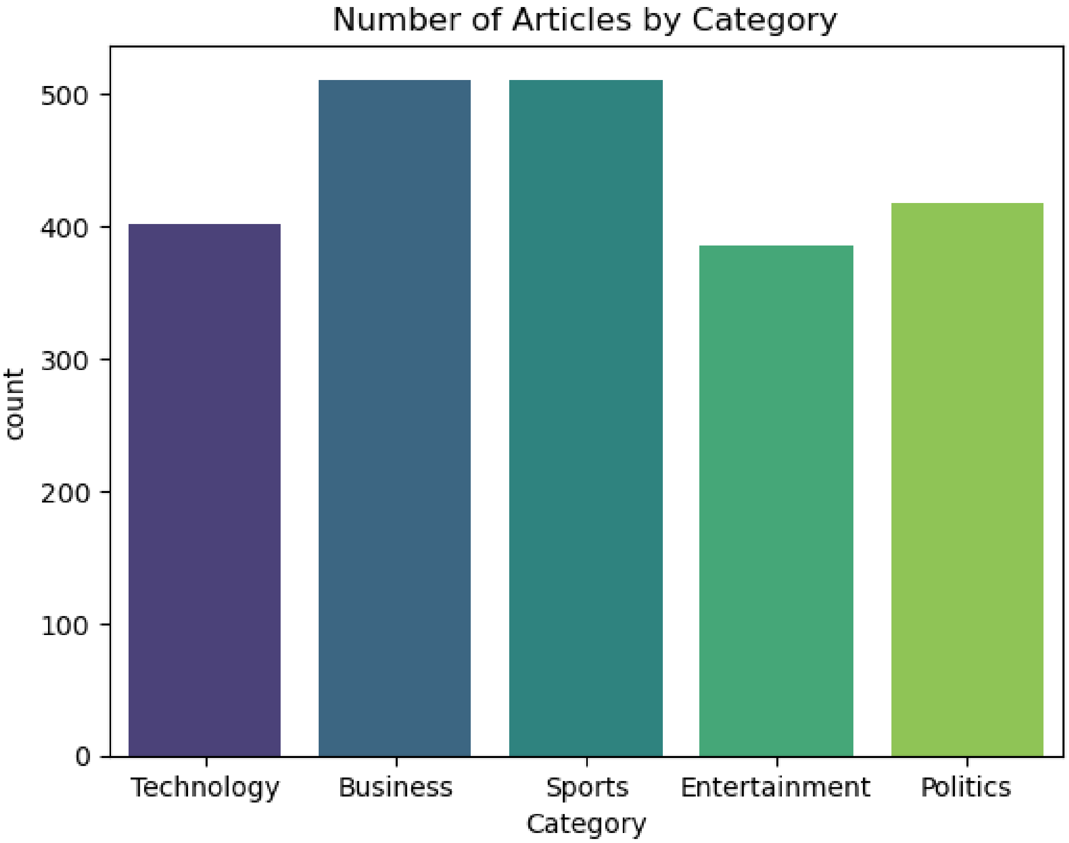
```
rf_model = RandomForestClassifier(random_state=42)
evaluate_model(rf_model, "Random Forest")
```

```
Random Forest Classification Report:
               precision    recall  f1-score   support

     Business       0.92      0.94      0.93       136
Entertainment       0.98      0.94      0.96        96
     Politics       0.93      0.96      0.94        98
       Sports       0.98      0.98      0.98       124
   Technology       0.97      0.94      0.96       103

     accuracy                           0.95       557
    macro avg       0.96      0.95      0.95       557
 weighted avg       0.95      0.95      0.95       557
```



[34]:

```
# Distribution of articles by category

sns.countplot(data=data, x='Category', palette='viridis')
plt.title('Number of Articles by Category')
plt.show()
```



14 Encoding and Transforming the data

Encoding the target variable # Label encoding of target variable

[36]:

```
label_encoder = LabelEncoder()
data['Category'] = label_encoder.fit_transform(data['Category'])

data.sample(10)
```

[36]:

```
      Category                                            Article  \
236          4  kenyan school turns to handhelds at the mbita …
```

```
2089        1  little britain vies for tv trophy bbc hits lit…
1314        4  ea to take on film and tv giants video game gi…
1095        3  african double in edinburgh world 5000m champi…
1744        2  brown to outline presidency goals next year wi…
338         0  aids and climate top davos agenda climate chan…
1650        2  brown s poll campaign move denied the governme…
1019        3  boro suffer morrison injury blow middlesbrough…
1398        4  beckham virus spotted on the net virus writers…
1761        3  campese berates whingeing england former austr…

                                      Processed_Article  Category_Label
236   kenyan school turn handhelds mbita point prima…               4
2089  little britain vies tv trophy bbc hit little b…               1
1314  ea take film tv giant video game giant electro…               4
1095  african double edinburgh world champion eliud …               3
1744  brown outline presidency goal next year make b…               2
338   aid climate top davos agenda climate change fi…               0
1650  brown poll campaign move denied government den…               2
1019  boro suffer morrison injury blow middlesbrough…               3
1398  beckham virus spotted net virus writer trading…               4
1761  campese berates whingeing england former austr…               3
```

15 Train-Test Split

16 Perform train-test split

[38]:
```
X = data[['Article']]
y = data['Category']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

[39]:
```
print(f'X_train shape: {X_train.shape}')
print(f'y_train shape: {y_train.shape}')
print(f'X_test shape: {X_test.shape}')
print(f'y_test shape: {y_test.shape}')
```

```
X_train shape: (1780, 1)
y_train shape: (1780,)
X_test shape: (445, 1)
y_test shape: (445,)
```

[45]:
```
Questionnaire:
Q1. How many news articles are present in the dataset that we have?
Ans1. 2225 articles are present in the dataset

Q2. Most of the news articles are from _____ category.
```

```
Ans2. Sports category has the maximum number (511) of articles

Q3. Only ___ no. of articles belong to the ‘Technology’ category.
Ans3. Only 401 articles belong to the technology class

Q4. What are Stop Words and why should they be removed from the text data?
Ans4. Stop words are words like I, you, and, because (pronouns, articles, conjunctions etc) which don't really add significance to the analysis of the document. They must be removed because they occur in high frequency but don't add value to the analysis.

Q5. Explain the difference between Stemming and Lemmatization.
Ans5. Stemming refers to simply truncating variations of the same words whereas lemmatization converts them to their actual root form. Lemmatization is more preferable because stemming words might result in words which don't even exist.

Q6. Which of the techniques Bag of Words or TF-IDF is considered to be more efficient than the other?
Ans6. The TF-IDF technique is more nuanced as it not only contains the information of normalised term frequency within a document (article), but also the number of documents the term appears in.

Q7. What’s the shape of train & test data sets after performing a 75:25 split. Ans7. The shape of train data is (1668,1) and test data is (557,1) after performing a 75:25 split.

Q8. Which of the following is found to be the best performing model..
a. Random Forest b. Nearest Neighbors c. Naive Bayes
Ans 8. Naive Bayes is the best performing model. Random Forest is also close in terms of performance.

Q9. According to this particular use case, both precision and recall are equally important. (T/F)
Ans9. True. Precision and Recall are both equally important. So F1 score is a better metric to look at.

Conclusion
The models which yielded the best accuracy of 96% were Naive Baye's Classifier and Random Forest Classifier
The different approaches to vectorize the data (BagofWords and TF-IDF) did not create much difference in the performance of models used - except for KNN Classifier. It was observed that KNN classifier yielded much higher accuracy with TF-IDF.
```

```
Overall, we were able to achieve a good accuracy even with a small dataset of␣
 ↪2225 articles as we were working on a balanced dataset with enough instances␣
 ↪of all classes for the model to train on.
```

```
  Cell In[45], line 8
    Q3. Only ___ no. of articles belong to the ‘Technology’ category.
                                               ^
SyntaxError: invalid character '‘' (U+2018)
```

```
[ ]:
```