In [3]:

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
from datasets import Dataset,DatasetDict
from transformers import AutoModelForSequenceClassification,AutoTokenizer
import statistics as st
from sklearn.cluster import AgglomerativeClustering, KMeans
from sklearn.metrics import accuracy_score, confusion_matrix
import time
import itertools
from sklearn.decomposition import NMF, LatentDirichletAllocation
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.svm import LinearSVC
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
```

```
None of PyTorch, TensorFlow >= 2.0, or Flax have been found. Models won't be a
vailable and only tokenizers, configuration and file/data utilities can be use
d.
```

In [24]:

```
# Load data and print a few rows
FILE_NAME = '/Users/wsyed2/Documents/JNTU/Python/Assignments/BBCNews.csv'

     # Loading the Data
df = pd.read_csv(FILE_NAME)

pd.options.display.max_rows
pd.set_option('display.max_colwidth', -1)
pd.set_option('display.width', 1000)
pd.set_option('display.max_columns', 50)
df.head()
```

Out[24]:

	ArticleId	
0	1833	worldcom ex-boss launches defence lawyers defending former worldcom chief bernie eb $11bn(£5.7bn)accountingfraud.mrebbershaspleadednotguiltytochargesoffraudandco$ $.prosecutionlawyershavearguedthatmrebbersorchestratedaseriesofaccountingtricks$ $.butmscooperwhonowrunsherownconsultingbusinesstoldajuryinnewyorkonwednesd$ $.mrebberslawyershavesaidhewasunawareofthefraudarguingthatauditorsdidnotaler$ $.mscooperalsosaidthatduringshareholdermeetingsmrebbersoftenpassedovertechnic$ $.theprosecutionsstarwitnessformerworldcomfinancialchiefscottsullivanhassaidth$ $.howevermscoopersaidmrsullivanhadnotmentionedanythinguncomfortableaboutwo$ $.worldcomemergedfrombankruptcyprotectionin2004andisnowknownasmci.lastweek$
1	154	german business confidence slides german business confidence fell in february knocking h the study found that the outlook in both the manufacturing and retail sectors had w economist bernd weidensteiner. the main reason is probably that the domestic economy is remained at a relatively high level and that he expected a modest economic upswing to co spend. latest indications are that growth is still proving elusive and ifo president hans-we record levels making german products less competitive overseas. on top of that the unemp that the ifo figures and germany s continuing problems may delay an interest ra
2	1101	bbc poll indicates economic gloom citizens in a majority of nations surveyed in a bbc world countries said they were positive about the future. almost 23 000 people in 22 countries we compared with respondents in nine countries who believed it was improving. those surveyed 48% were pessimistic about their national economy while 41% were optimistic. and 47% saw program on international policy attitudes (pipa) at the university of maryland. while the world e kull. people around the world are saying: i m ok but the world isn t . there may be a percept blunt. the countries where people were most optimistic both for the world and for their ow scale says the bbc s louisa lim in beijing. but the results also may reflect the untrammelled c in italy and mexico were also quite gloomy. the bbc s david willey in rome says one reaso among the most upbeat countries on prospects for respondents families but one of the be
3	1976	lifestyle governs mobile choice faster better or funkier hardware alone is not going to hel interested in how handsets fit in with their lifestyle than they are in screen size onboard mobile media at ericsson s consumer and enterprise lab. we have to stop saying that these study ericsson interviewed 14 000 mobile phone owners on the ways they use their phone. pe people he said. while diaries have always been popular a mobile phone -- especially one slightly changed way. dr bjorn said that although consumers do what they always did bu different tribes that use phones in different ways. dr bjorn said groups dubbed pioneers and this was because younger users often the children of ageing mobile owners encou materialists. only when about 25% of people have handsets with new innovations or innovations tends to take off. dr bjorn said that early reports of camera phone usage in japa was 29%. similarly across europe the numbers of people taking snaps with cameras is starti people also used their camera phones in very different ways to film and even digital camer
4	917	enron bosses in $168mpayouteighteenformerenrondirectorshaveagreeda168$ $13mfromtheirownpockets.thesettlementwillbeputtothecourtsforapprovalnextweek.$ $.beforeitscollapsethefirmwastheseventhbiggestpublicuscompanybyrevenue.itsdemi$ $.thesettlementisverysignificantinholdingtheseoutsidedirectorsatleastpartiallyperso$ $.hopefullythiswillhelpsendamessagetocorporateboardroomsoftheimportanceofdire$ $155mofwhichwillbecoveredbyinsurance-noneofthe18formerdirectorswilladmitan$ $.sofarincludingthelatestdealjustunder$ 500m (£378.8m) has been retrieved for investors. however the latest deal does not include has pleaded guilty to taking part in an illegal conspiracy while he was chief financial officer first boston. the university of california said the

In [15]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1490 entries, 0 to 1489
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   ArticleId  1490 non-null   int64 
 1   Text       1490 non-null   object
 2   Category   1490 non-null   object
dtypes: int64(1), object(2)
memory usage: 35.0+ KB
```

In [26]:

```
print("Total observations ", len(df))
print("Total Count of Unique Article IDs ", len(df['ArticleId'].unique()))
```

```
Total observations  1490
Total Count of Unique Article IDs  1490
```

In [19]:

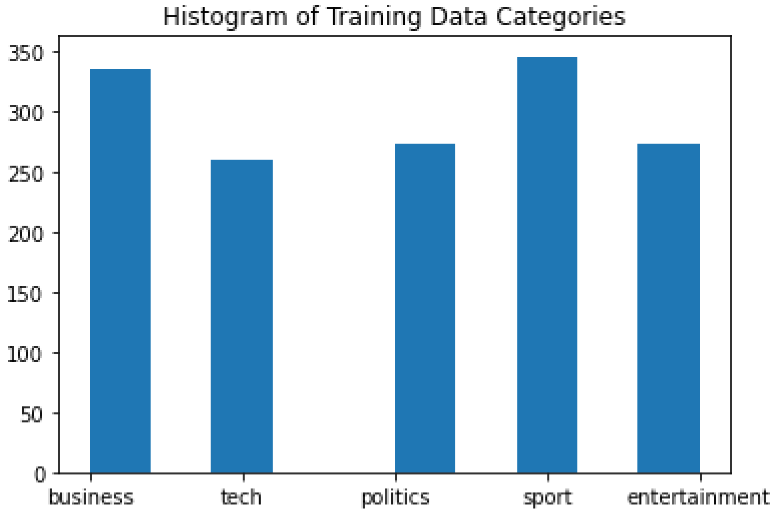
Out[19]:

	Category	categoryId
0	business	0
3	tech	1
5	politics	2
6	sport	3
7	entertainment	4

In [28]:

```
plt.hist(df['Category'])
plt.title("Histogram of Training Data Categories")
print("Tech the smallest category makes up this percentage:", round(len(df[df['
print("Sport the largest category makes up this percentage:", round(len(df[df['
categories = df['Category'].unique()
```

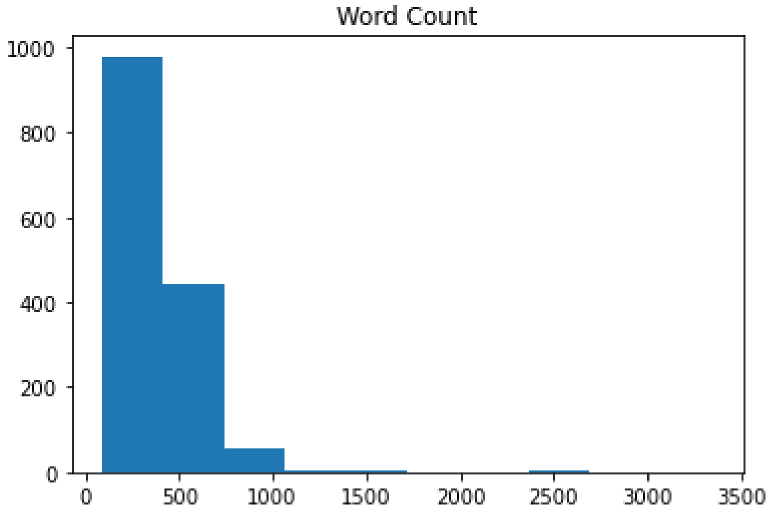
```
Tech the smallest category makes up this percentage: 0.175
Sport the largest category makes up this percentage: 0.232
```



In [30]:

```
a = []
for txt in df['Text']:
    a.append(len(txt.split()))
plt.hist(a)
plt.title("Word Count")
print("Smallest Article ", min(a))
print("Largest Article " , max(a))
```

```
Smallest Article  90
Largest Article  3345
```



In [32]:
```
no_features = 2000
text_ds = df['Text']
# NMF is able to use tf-idf
tfidf_vectorizer = TfidfVectorizer(max_df=0.95, min_df=2, max_features=no_featu
tfidf = tfidf_vectorizer.fit_transform(text_ds)
tfidf_feature_names = tfidf_vectorizer.get_feature_names()
```

In [33]:
```
no_topics = 5

# Run NMF
nmf = NMF(no_topics, random_state=1, alpha=.1, l1_ratio=.5, init='nndsvd').fit(
```

In [34]:
```
def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print ("Topic %d:" % (topic_idx))
        print (" ".join([feature_names[i]
                        for i in topic.argsort()[:-no_top_words - 1:-1]]))

no_top_words = 10

display_topics(nmf, tfidf_feature_names, no_top_words)
```

```
Topic 0:
england game win said wales cup ireland play players team
Topic 1:
mr labour blair election brown party said government minister prime
Topic 2:
mobile people music said phone technology digital users phones software
Topic 3:
film best awards award actor oscar actress films festival director
Topic 4:
said growth economy year sales market bank oil economic 2004
```

In [ ]:

In [ ]: