Assignment 12

1. Text Classification of News Articles using NLP.

```
Article Id – Article id unique given to the record
Article – Text of the header and article
Category – Category of the article (tech, business, sport, entertainment, politics)
```

Consider BBC News as corpus for implementing question 1

In [194]:
```
#import important libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import os

#EDA and preprocessing
import re
import nltk.corpus
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from string import digits

#modeling
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.decomposition import NMF
from sklearn.metrics import accuracy_score
import sklearn.metrics as metrics
import itertools
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split

import warnings
warnings.filterwarnings('ignore')
```

In [195]:
```
file_path = 'Documents/Data science Material JNTU/Assignments/Assignment 12/BBC News.csv'
```

In [196]:
```
data = pd.read_csv(file_path)
```

Exploratory Data Analysis

In [197]:
```
data
```

Out[197]:

	ArticleId	Text	Category
0	1833	worldcom ex-boss launches defence lawyers defe...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll indicates economic gloom citizens in ...	business
3	1976	lifestyle governs mobile choice faster bett...	tech
4	917	enron bosses in $168m payout eighteen former e...	business
...	...	...	...
1485	857	double eviction from big brother model caprice...	entertainment
1486	325	dj double act revamp chart show dj duo jk and ...	entertainment
1487	1590	weak dollar hits reuters revenues at media gro...	business
1488	1587	apple ipod family expands market apple has exp...	tech
1489	538	santy worm makes unwelcome visit thousands of ...	tech

1490 rows × 3 columns

In [198]:
```
data.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 [199]:
```
data['ArticleId'].nunique()
```

Out[199]:
```
1490
```

In [200]:
```
data['Category'].unique()
```

Out[200]:
```
array(['business', 'tech', 'politics', 'sport', 'entertainment'],
      dtype=object)
```

In [201]:
```
data['Category'].value_counts()
```

Out[201]:
```
sport            346
business         336
politics         274
entertainment    273
tech             261
Name: Category, dtype: int64
```

In [202]:
```
data['Category'].value_counts().sum()
```

Out[202]:
```
1490
```

In [203]:
```
data['Text'].value_counts().sum()
```

Out[203]:
```
1490
```

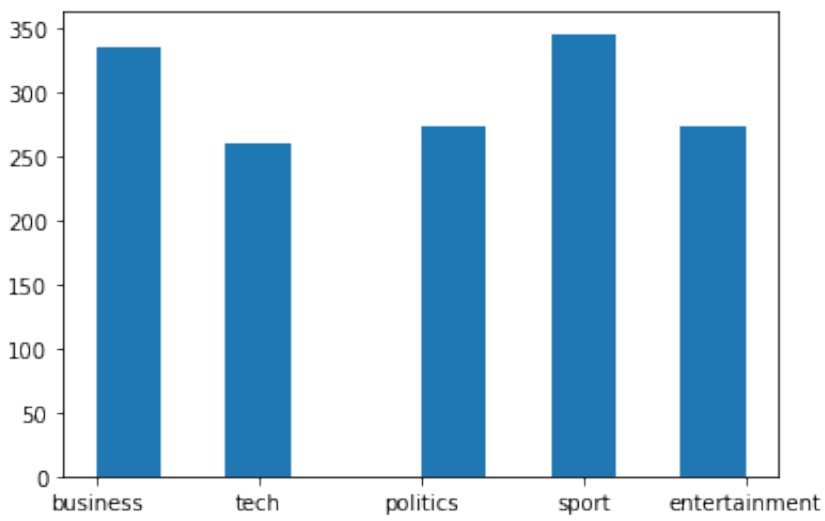
In [204]:
```
data.head(5)
```

Out[204]:

	ArticleId	Text	Category
0	1833	worldcom ex-boss launches defence lawyers defe...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll indicates economic gloom citizens in ...	business
3	1976	lifestyle governs mobile choice faster bett...	tech
4	917	enron bosses in $168m payout eighteen former e...	business

In [205]:
```
plt.hist(data['Category'])
```

Out[205]:
```
(array([336.,   0., 261.,   0.,   0., 274.,   0., 346.,   0., 273.]),
 array([0. , 0.4, 0.8, 1.2, 1.6, 2. , 2.4, 2.8, 3.2, 3.6, 4. ]),
 <BarContainer object of 10 artists>)
```



In [206]:

```
#checking some random text data to verify are there any unneccessary data present

print(data['Text'][0])
print('____________________')
print(data['Text'][500])
print('____________________')
print(data['Text'][1000])
```

worldcom ex-boss launches defence lawyers defending former worldcom chief bernie ebbers against a batte
ry of fraud charges have called a company whistleblower as their first witness. cynthia cooper worldc
om s ex-head of internal accounting alerted directors to irregular accounting practices at the us tele
coms giant in 2002. her warnings led to the collapse of the firm following the discovery of an $11bn (£
5.7bn) accounting fraud. mr ebbers has pleaded not guilty to charges of fraud and conspiracy. prosecut
ion lawyers have argued that mr ebbers orchestrated a series of accounting tricks at worldcom ordering
employees to hide expenses and inflate revenues to meet wall street earnings estimates. but ms cooper
who now runs her own consulting business told a jury in new york on wednesday that external auditors a
rthur andersen had approved worldcom s accounting in early 2001 and 2002. she said andersen had given a
green light to the procedures and practices used by worldcom. mr ebber s lawyers have said he was unaw
are of the fraud arguing that auditors did not alert him to any problems. ms cooper also said that du
ring shareholder meetings mr ebbers often passed over technical questions to the company s finance chie
f giving only brief answers himself. the prosecution s star witness former worldcom financial chief
scott sullivan has said that mr ebbers ordered accounting adjustments at the firm telling him to hit
our books . however ms cooper said mr sullivan had not mentioned anything uncomfortable about worldc
om s accounting during a 2001 audit committee meeting. mr ebbers could face a jail sentence of 85 years
if convicted of all the charges he is facing. worldcom emerged from bankruptcy protection in 2004 and
is now known as mci. last week mci agreed to a buyout by verizon communications in a deal valued at $6
.75bn.

____

gm in crunch talks on fiat future fiat will meet car giant general motors (gm) on tuesday in an attempt
to reach agreement over the future of the italian firm s loss-making auto group. fiat claims that gm i
s legally obliged to buy the 90% of the car unit it does not already own; gm says the contract signed
in 2000 is no longer valid. press reports have speculated that fiat may be willing to accept a cash pa
yment in return for dropping its claim. both companies want to cut costs as the car industry adjusts to
waning demand. the meeting between fiat boss sergio marchionne and gm s rick wagoner is due to take pl
ace at 1330 gmt in zurich according to the reuters news agency. mr marchionne is confident of his fir
m s legal position saying in an interview with the financial times that gm s argument has no legs . t
he agreement in question dates back to gm s decision to buy 20% of fiat s auto division in 2000. at the
time it gave the italian firm the right via a put option to sell the remaining stake to gm. in rec
ent weeks fiat has reiterated its claims that this put is still valid and legally binding. however
gm argues that a fiat share sale made last year which cut gm s holding to 10% together with asset sal
es made by fiat have terminated the agreement. selling the fiat s car-making unit may not prove so sim
ple analysts say especially as it is a company that is so closely linked to italy s industrial herita
ge. political and public pressure may well push the two firms to reach a compromise. we are not expect
ing fiat to exercise its put of the auto business against an unwilling gm at this point brokerage mer
rill lynch said in a note to investors adding that any legal battle would be protracted and damaging t
o the business. as far as we are aware the agnelli family which indirectly controls at least 30% of
fiat has not given a firm public indication that it wants to sell the auto business. fiat may be will
ing to cancel the put in exchange for money.

____

middlesbrough 2-2 charlton a late header by teenager danny graham earned middlesbrough a battling draw
with charlton at the riverside. matt holland had put the visitors ahead in the 14th minute after his s
hot took a deflection off franck queudrue. but middlesbrough peppered the charlton goal after the break
and chris riggott stroked home the equaliser. shaun bartlett s strike put charlton back in front but th
at lead lasted just six minutes before graham rushed onto queudrue s pass to head home. the match burst
to life from the whistle and charlton defender hermann hreidarsson had sight of an open goal after just
six minutes. hreidarsson received danny murphy s free-kick from the right but he crashed his free heade
r wide of the far post. the iceland international looked such a danger the boro bench could be heard is
suing frantic instructions to mark him. charlton s early pressure paid off when bartlett received a lo
ng ball from talal el karkouri in the box and laid it off to holland who buried his right-footed strike
. szilard nemeth recalled in place of joseph-desire job was twice denied his chance to get middlesbro
ugh back on level terms by dean kiely. the striker played a great one-two with jimmy floyd hasselbaink
only to see kiely get down well to smother his shot before directing a header straight into the keeper
s arms. boro had plenty of time on the ball but the addicks comfortably mopped up the pressure - with k
iely tipping a hasselbaink header over the bar - to take their lead into half-time. it was all one-way
traffic after the break at the riverside as middlesbrough poured forward and kiely even saved hreidarss
on s blushes when he palmed the ball away to prevent a charlton own goal. but the addicks keeper could
do nothing about riggott s equaliser in the 74th minute. the boro defender looked suspiciously offside
as he got on the end of gareth southgate s misdirected effort but despite the charlton protests his go
al stood. the addicks did not let their heads drop and bartlett left the boro defence standing picking
up hreidarsson s cross to easily sink his right-footed strike. but substitute graham was on hand to gra
b a share of the points for the home side. the 19-year-old striker nodding home the equaliser - and his
first premiership goal - with five minutes left on the clock. i felt we did enough to win the game ev
en though the first half was lacklustre. we dominated after the break the players showed a fantastic
response and we should have gone on to win. but for (charlton goalkeeper) dean kiely who made three t
remendous saves we could have scored five or six. to take the lead and then to get penned back it
feels a little bit like a defeat admitted kiely. we were winning but middlesbrough kept knocking on
the door. but we stood up and credit to us we didn t capitulate. we ll kick on now. our short-term amb
ition is to progress from the seventh place finish from last year. nash reiziger (graham 82) riggot
t southgate queudrue parlour (job 86) doriva nemeth (parnaby 87) zenden downing hasselbaink. s
ubs not used: cooper knight. riggott 74 graham 86. kiely hreidarsson perry el karkouri young k
onchesky murphy (euell 78) holland kishishev thomas (johansson 72) bartlett. subs not used: fish
jeffers andersen. konchesky hreidarsson perry. holland 14 bartlett 80. 29 603 m riley (w yorksh
ire).

As you can clearly see there are many un wanted text in between the data. like numbers, special character, spaces and stop words. So we need to remove unneccessary data from the text.

Removing all punctuations

In [207]:
```
data['without_punct'] = data['Text'].apply(lambda row: re.sub(r'[^\w\s]+', '', row))
```

In [208]:
```
data['without_punct'][0]
```

Out[208]:
```
'worldcom exboss launches defence lawyers defending former worldcom chief bernie ebbers against a batte
ry of fraud charges have called a company whistleblower as their first witness  cynthia cooper  worldco
m s exhead of internal accounting  alerted directors to irregular accounting practices at the us teleco
ms giant in 2002 her warnings led to the collapse of the firm following the discovery of an 11bn 57bn a
ccounting fraud mr ebbers has pleaded not guilty to charges of fraud and conspiracy  prosecution lawyer
s have argued that mr ebbers orchestrated a series of accounting tricks at worldcom  ordering employees
to hide expenses and inflate revenues to meet wall street earnings estimates but ms cooper  who now run
s her own consulting business  told a jury in new york on wednesday that external auditors arthur ander
sen had approved worldcom s accounting in early 2001 and 2002 she said andersen had given a  green ligh
t  to the procedures and practices used by worldcom mr ebber s lawyers have said he was unaware of the
fraud  arguing that auditors did not alert him to any problems  ms cooper also said that during shareho
lder meetings mr ebbers often passed over technical questions to the company s finance chief  giving on
ly  brief  answers himself the prosecution s star witness  former worldcom financial chief scott sulliv
an  has said that mr ebbers ordered accounting adjustments at the firm  telling him to  hit our books
however  ms cooper said mr sullivan had not mentioned  anything uncomfortable  about worldcom s account
ing during a 2001 audit committee meeting mr ebbers could face a jail sentence of 85 years if convicted
of all the charges he is facing worldcom emerged from bankruptcy protection in 2004  and is now known a
s mci last week  mci agreed to a buyout by verizon communications in a deal valued at 675bn'
```

Removing numbers

In [209]:
```
data['without_punct_num'] = data['without_punct'].apply(lambda row: re.sub(r'[0-9]+', '', row))
```

In [210]:
```
data['without_punct_num'][0]
```

Out[210]:
```
'worldcom exboss launches defence lawyers defending former worldcom chief bernie ebbers against a batte
ry of fraud charges have called a company whistleblower as their first witness  cynthia cooper  worldco
m s exhead of internal accounting  alerted directors to irregular accounting practices at the us teleco
ms giant in  her warnings led to the collapse of the firm following the discovery of an bn bn accountin
g fraud mr ebbers has pleaded not guilty to charges of fraud and conspiracy  prosecution lawyers have a
rgued that mr ebbers orchestrated a series of accounting tricks at worldcom  ordering employees to hide
expenses and inflate revenues to meet wall street earnings estimates but ms cooper  who now runs her ow
n consulting business  told a jury in new york on wednesday that external auditors arthur andersen had
approved worldcom s accounting in early  and  she said andersen had given a  green light  to the proced
ures and practices used by worldcom mr ebber s lawyers have said he was unaware of the fraud  arguing t
hat auditors did not alert him to any problems  ms cooper also said that during shareholder meetings mr
ebbers often passed over technical questions to the company s finance chief  giving only  brief  answer
s himself the prosecution s star witness  former worldcom financial chief scott sullivan  has said that
mr ebbers ordered accounting adjustments at the firm  telling him to  hit our books  however  ms cooper
said mr sullivan had not mentioned  anything uncomfortable  about worldcom s accounting during a  audit
committee meeting mr ebbers could face a jail sentence of  years if convicted of all the charges he is
facing worldcom emerged from bankruptcy protection in  and is now known as mci last week  mci agreed t
o a buyout by verizon communications in a deal valued at bn'
```

Removing stopwords

In [211]:
```
op_words = stopwords.words('english')
ata['without_stopwords'] = data['without_punct_num'].apply(lambda x: ' '.join([word for word in x.split()
```

In [212]:
```
data['without_stopwords'][0]
```

Out[212]:
```
'worldcom exboss launches defence lawyers defending former worldcom chief bernie ebbers battery fraud c
harges called company whistleblower first witness cynthia cooper worldcom exhead internal accounting al
erted directors irregular accounting practices us telecoms giant warnings led collapse firm following d
iscovery bn bn accounting fraud mr ebbers pleaded guilty charges fraud conspiracy prosecution lawyers a
rgued mr ebbers orchestrated series accounting tricks worldcom ordering employees hide expenses inflate
revenues meet wall street earnings estimates ms cooper runs consulting business told jury new york wedn
esday external auditors arthur andersen approved worldcom accounting early said andersen given green li
ght procedures practices used worldcom mr ebber lawyers said unaware fraud arguing auditors alert probl
ems ms cooper also said shareholder meetings mr ebbers often passed technical questions company finance
chief giving brief answers prosecution star witness former worldcom financial chief scott sullivan said
mr ebbers ordered accounting adjustments firm telling hit books however ms cooper said mr sullivan ment
ioned anything uncomfortable worldcom accounting audit committee meeting mr ebbers could face jail sent
ence years convicted charges facing worldcom emerged bankruptcy protection known mci last week mci agre
ed buyout verizon communications deal valued bn'
```

Removing Extra Spaces between words

In [213]:
```
data['Final_Text'] = data['without_stopwords'].apply(lambda x: re.sub('\s+', ' ', x))
```

In [214]:
```
data['Final_Text'][0]
```

Out[214]:
```
'worldcom exboss launches defence lawyers defending former worldcom chief bernie ebbers battery fraud c
harges called company whistleblower first witness cynthia cooper worldcom exhead internal accounting al
erted directors irregular accounting practices us telecoms giant warnings led collapse firm following d
iscovery bn bn accounting fraud mr ebbers pleaded guilty charges fraud conspiracy prosecution lawyers a
rgued mr ebbers orchestrated series accounting tricks worldcom ordering employees hide expenses inflate
revenues meet wall street earnings estimates ms cooper runs consulting business told jury new york wedn
esday external auditors arthur andersen approved worldcom accounting early said andersen given green li
ght procedures practices used worldcom mr ebber lawyers said unaware fraud arguing auditors alert probl
ems ms cooper also said shareholder meetings mr ebbers often passed technical questions company finance
chief giving brief answers prosecution star witness former worldcom financial chief scott sullivan said
mr ebbers ordered accounting adjustments firm telling hit books however ms cooper said mr sullivan ment
ioned anything uncomfortable worldcom accounting audit committee meeting mr ebbers could face jail sent
ence years convicted charges facing worldcom emerged bankruptcy protection known mci last week mci agre
ed buyout verizon communications deal valued bn'
```

Word Tokenizer

In [215]:
```
data['tokenized'] = data.apply(lambda row: nltk.word_tokenize(row['Final_Text']), axis=1)
```

In [216]:
```
data['tokenized'][0]
```

Out[216]:
```
['worldcom',
 'exboss',
 'launches',
 'defence',
 'lawyers',
 'defending',
 'former',
 'worldcom',
 'chief',
 'bernie',
 'ebbers',
 'battery',
 'fraud',
 'charges',
 'called',
 'company',
 'whistleblower',
 'first',
 'witness',
 'cynthia',
```

Lemmetization

In [217]:
```
wordnet_lemmatizer = WordNetLemmatizer()

def lemmatizer(text):
    lem = [wordnet_lemmatizer.lemmatize(word.lower()) for word in text]
    return lem

data['lemmatized'] = data['tokenized'].apply(lambda string: lemmatizer(string))
```

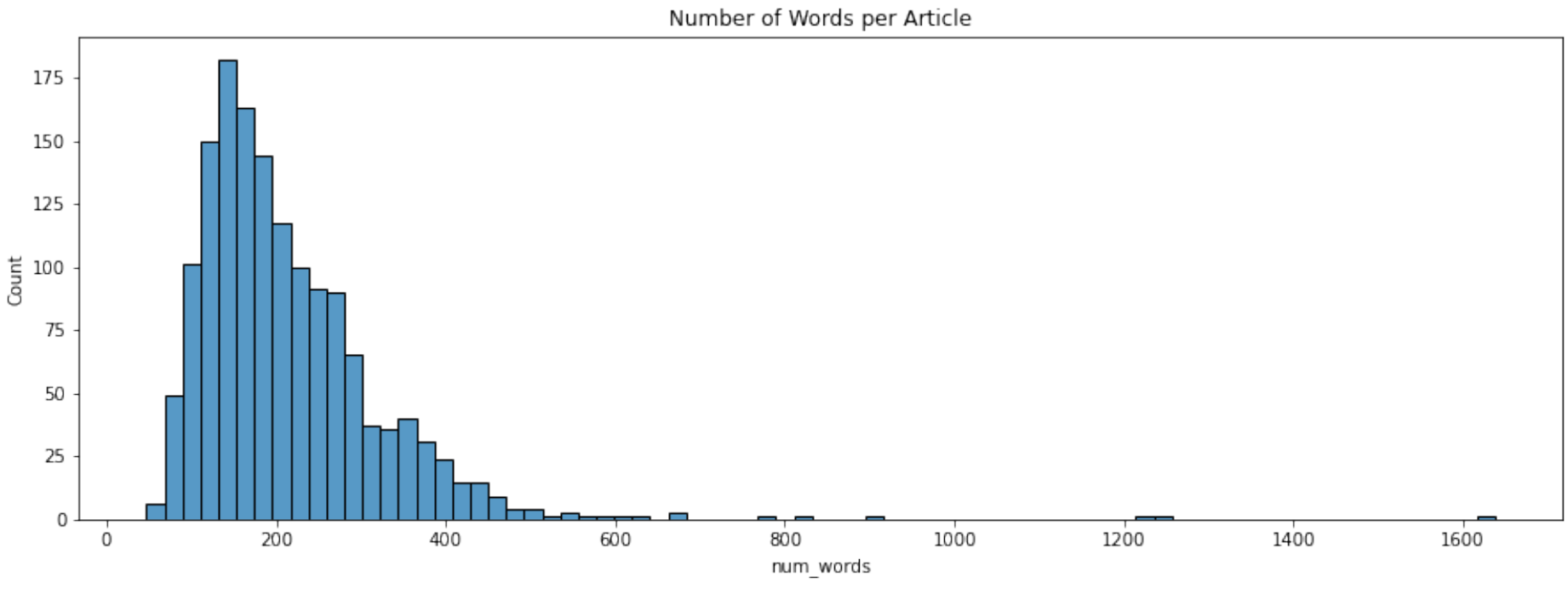
In [218]:
```
data['lemmatized'][0]
```

```
 'year',
 'convicted',
 'charge',
 'facing',
 'worldcom',
 'emerged',
 'bankruptcy',
 'protection',
 'known',
 'mci',
 'last',
 'week',
 'mci',
 'agreed',
 'buyout',
 'verizon',
 'communication',
 'deal',
 'valued',
 'bn']
```

Counting Number of words in each article

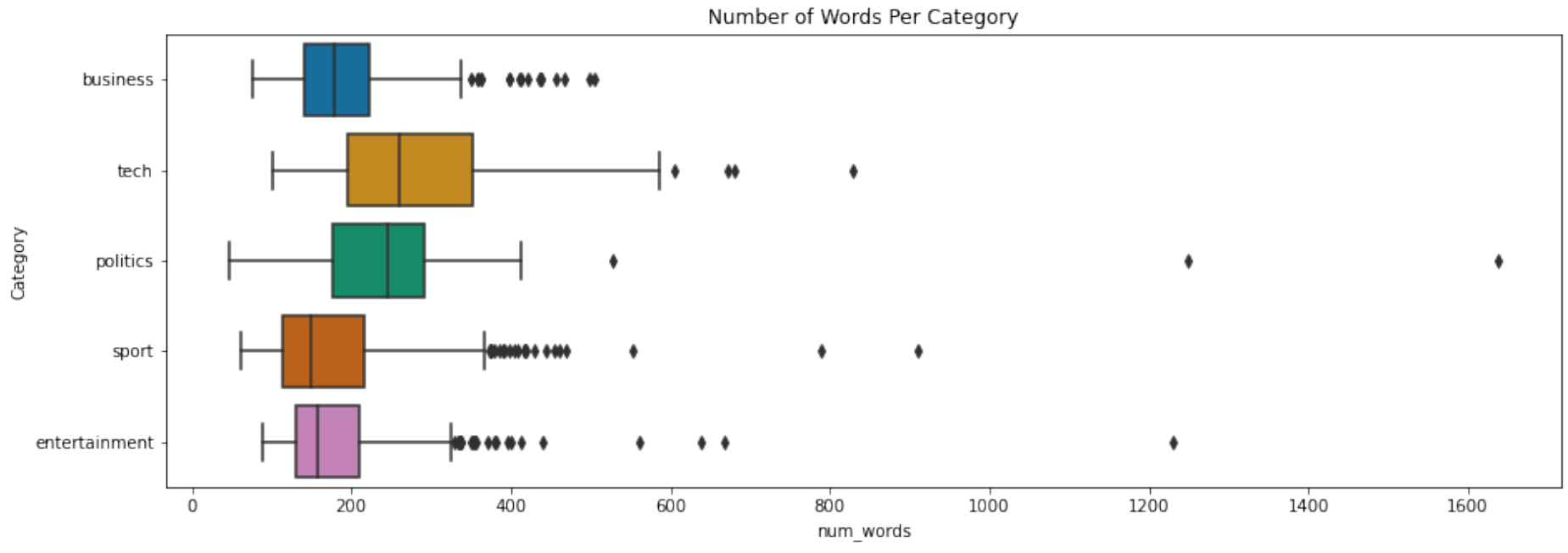
In [219]:
```
data['num_words'] = data['lemmatized'].apply(lambda lst: len(lst))
```

In [220]:
```
# number of tokens (words) per article
fig, ax = plt.subplots(figsize=(15, 5))
sns.histplot(
    data = data,
    x = 'num_words',
    palette = 'colorblind',
    ).set(
        title = 'Number of Words per Article');
```



Outlier Analysis

In [221]:
```
# words per category
fig, ax = plt.subplots(figsize=(15, 5))
sns.boxplot(
    data = data,
    x = 'num_words',
    y = 'Category',
    palette = 'colorblind'
    ).set(
        title = 'Number of Words Per Category');
```



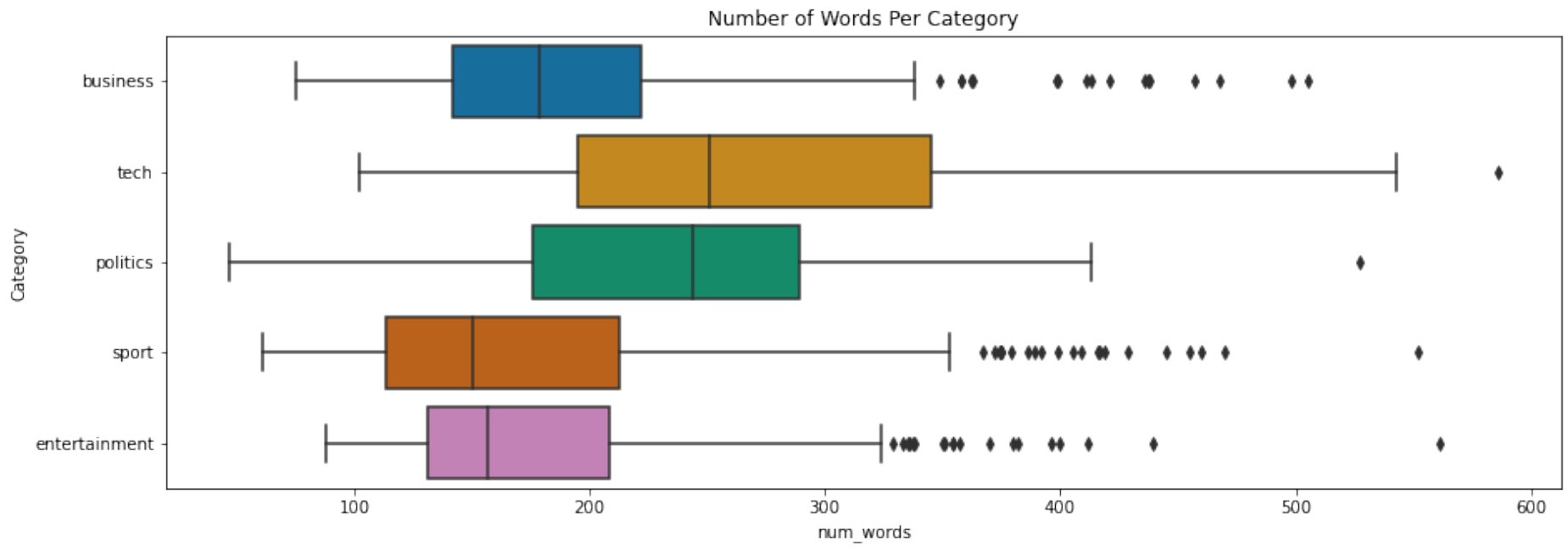
As we can see there are outliers. So, we need to remove these outliers before building the model. By seeing histogram graph we can consider 600 as the boundary to delete outliers.

In [222]:
```
#remove outlier texts longer than 600 words

data = data[data['num_words'] < 600]
len(data)
```

Out[222]: 1479

In [223]:
```
# words per category
fig, ax = plt.subplots(figsize=(15, 5))
sns.boxplot(
    data = data,
    x = 'num_words',
    y = 'Category',
    palette = 'colorblind'
    ).set(
        title = 'Number of Words Per Category');
```



As you can see now outliers are greatly reduced. Now we can go ahead with the model building.

Model Building

1. Naïve Bayes

2. Linear SVC

In [224]:
```
from sklearn.model_selection import train_test_split

X = data['Final_Text']
y = data['Category']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

In [225]:
```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC

# Naïve Bayes:
text_clf_nb = Pipeline([('tfidf', TfidfVectorizer()),
                        ('clf', MultinomialNB()),
])

# Linear SVC:
text_clf_lsvc = Pipeline([('tfidf', TfidfVectorizer()),
                        ('clf', LinearSVC()),
])
```

In [226]:
```
text_clf_nb.fit(X_train, y_train)
```

Out[226]:
```
Pipeline(steps=[('tfidf', TfidfVectorizer()), ('clf', MultinomialNB())])
```

In [227]:
```
# Form a prediction set
predictions = text_clf_nb.predict(X_test)
```

In [228]:
```
# Report the confusion matrix
from sklearn import metrics
print(metrics.confusion_matrix(y_test,predictions))
```

```
[[107   0   0   0   0]
 [  3  89   4   2   2]
 [  3   0  79   0   0]
 [  0   0   0 116   0]
 [  1   0   1   1  81]]
```

In [229]:
```
# Print a classification report
print(metrics.classification_report(y_test,predictions))
```

```
               precision    recall  f1-score   support

     business       0.94      1.00      0.97       107
entertainment       1.00      0.89      0.94       100
     politics       0.94      0.96      0.95        82
        sport       0.97      1.00      0.99       116
         tech       0.98      0.96      0.97        84

     accuracy                           0.97       489
    macro avg       0.97      0.96      0.96       489
 weighted avg       0.97      0.97      0.96       489
```

In [230]:
```
# Printing the overall accuracy of Naive bayes
print(metrics.accuracy_score(y_test,predictions))
```

```
0.9652351738241309
```

In [231]:
```
#Model fitting using Linear SVC
text_clf_lsvc.fit(X_train, y_train)
```

Out[231]:
```
Pipeline(steps=[('tfidf', TfidfVectorizer()), ('clf', LinearSVC())])
```

In [232]:
```
# Form a prediction set
predictions = text_clf_lsvc.predict(X_test)
```

In [233]:
```
# Report the confusion matrix
from sklearn import metrics
print(metrics.confusion_matrix(y_test,predictions))
```

```
[[105   0   1   0   1]
 [  1  95   3   0   1]
 [  3   0  79   0   0]
 [  0   0   0 116   0]
 [  0   1   0   0  83]]
```

In [234]:
```
# Print a classification report
print(metrics.classification_report(y_test,predictions))
```

```
                precision    recall  f1-score   support

      business       0.96      0.98      0.97       107
 entertainment       0.99      0.95      0.97       100
      politics       0.95      0.96      0.96        82
         sport       1.00      1.00      1.00       116
          tech       0.98      0.99      0.98        84

      accuracy                           0.98       489
     macro avg       0.98      0.98      0.98       489
  weighted avg       0.98      0.98      0.98       489
```

In [235]:
```
# Printing the overall accuracy of Linear SVC
print(metrics.accuracy_score(y_test,predictions))
```

```
0.9775051124744376
```

Conclusion

As you can clearly see both Naive Bayes & Linear SVC have performed very well.

Naive Bayes Accuaracy is - 96.5%

Linear SVC Accuracy is - 97.7 %