5222: Feature Engineering ICE#1

Rubric

1. Complete and proper Github Submission (10%)
2. Complete and proper submission to Canvas (5%)
3. Source Code (50%)
4. Explaining the answers (30%)
5. Commenting, formatting, and visualizing your code properly and timely submission (5%)

Please goto https://towardsdatascience.com/text-classification-in-python-dd95d264c802 (https://towardsdatascience.com/text-classification-in-python-dd95d264c802) and follow the article.

This article is for text classification using python.

Their Github is available at https://github.com/miguelfzafra/Latest-News-Classifier/tree/master/0.%20Latest%20News%20Classifier (https://github.com/miguelfzafra/Latest-News-Classifier/tree/master/0.%20Latest%20News%20Classifier).

Follow their step 00,01,02,03 and 04. Use the same workbook to execute your code

In [1]:

```
!pip install altair vega_datasets notebook vega
import pandas as pd
import matplotlib.pyplot as plt
import pickle
import seaborn as sns
sns.set_style("whitegrid")
import altair as alt
alt.renderers.enable("notebook")

# Code for hiding seaborn warnings
import warnings
warnings.filterwarnings("ignore")
```

```
Requirement already satisfied: altair in c:\users\deeks\anaconda3\lib\site-
packages (4.2.0)
Requirement already satisfied: vega_datasets in c:\users\deeks\anaconda3\li
b\site-packages (0.9.0)
Requirement already satisfied: notebook in c:\users\deeks\anaconda3\lib\sit
e-packages (6.4.5)
Requirement already satisfied: vega in c:\users\deeks\anaconda3\lib\site-pa
ckages (3.6.0)
Requirement already satisfied: entrypoints in c:\users\deeks\anaconda3\lib
\site-packages (from altair) (0.3)
Requirement already satisfied: jsonschema>=3.0 in c:\users\deeks\anaconda3
\lib\site-packages (from altair) (3.2.0)
Requirement already satisfied: numpy in c:\users\deeks\anaconda3\lib\site-p
ackages (from altair) (1.20.3)
Requirement already satisfied: pandas>=0.18 in c:\users\deeks\anaconda3\lib
\site-packages (from altair) (1.3.4)
Requirement already satisfied: jinja2 in c:\users\deeks\anaconda3\lib\site-
packages (from altair) (2.11.3)
Requirement already satisfied: toolz in c:\users\deeks\anaconda3\lib\site-p
ackages (from altair) (0.11.1)
Requirement already satisfied: tornado>=6.1 in c:\users\deeks\anaconda3\lib
\site-packages (from notebook) (6.1)
Requirement already satisfied: pyzmq>=17 in c:\users\deeks\anaconda3\lib\si
te-packages (from notebook) (22.2.1)
Requirement already satisfied: jupyter-core>=4.6.1 in c:\users\deeks\anacon
da3\lib\site-packages (from notebook) (4.8.1)
Requirement already satisfied: ipykernel in c:\users\deeks\anaconda3\lib\si
te-packages (from notebook) (6.4.1)
Requirement already satisfied: Send2Trash>=1.5.0 in c:\users\deeks\anaconda
3\lib\site-packages (from notebook) (1.8.0)
Requirement already satisfied: traitlets>=4.2.1 in c:\users\deeks\anaconda3
\lib\site-packages (from notebook) (5.1.0)
Requirement already satisfied: ipython-genutils in c:\users\deeks\anaconda3
\lib\site-packages (from notebook) (0.2.0)
Requirement already satisfied: argon2-cffi in c:\users\deeks\anaconda3\lib
\site-packages (from notebook) (20.1.0)
Requirement already satisfied: prometheus-client in c:\users\deeks\anaconda
3\lib\site-packages (from notebook) (0.11.0)
Requirement already satisfied: nbformat in c:\users\deeks\anaconda3\lib\sit
e-packages (from notebook) (5.1.3)
Requirement already satisfied: nbconvert in c:\users\deeks\anaconda3\lib\si
te-packages (from notebook) (6.1.0)
Requirement already satisfied: terminado>=0.8.3 in c:\users\deeks\anaconda3
```

```
\lib\site-packages (from notebook) (0.9.4)
Requirement already satisfied: jupyter-client>=5.3.4 in c:\users\deeks\anac
onda3\lib\site-packages (from notebook) (6.1.12)
Requirement already satisfied: jupyter<2.0.0,>=1.0.0 in c:\users\deeks\anac
onda3\lib\site-packages (from vega) (1.0.0)
Requirement already satisfied: pyrsistent>=0.14.0 in c:\users\deeks\anacond
a3\lib\site-packages (from jsonschema>=3.0->altair) (0.18.0)
Requirement already satisfied: attrs>=17.4.0 in c:\users\deeks\anaconda3\li
b\site-packages (from jsonschema>=3.0->altair) (21.2.0)
Requirement already satisfied: six>=1.11.0 in c:\users\deeks\anaconda3\lib
\site-packages (from jsonschema>=3.0->altair) (1.16.0)
Requirement already satisfied: setuptools in c:\users\deeks\anaconda3\lib\s
ite-packages (from jsonschema>=3.0->altair) (58.0.4)
Requirement already satisfied: qtconsole in c:\users\deeks\anaconda3\lib\si
te-packages (from jupyter<2.0.0,>=1.0.0->vega) (5.1.1)
Requirement already satisfied: ipywidgets in c:\users\deeks\anaconda3\lib\s
ite-packages (from jupyter<2.0.0,>=1.0.0->vega) (7.6.5)
Requirement already satisfied: jupyter-console in c:\users\deeks\anaconda3
\lib\site-packages (from jupyter<2.0.0,>=1.0.0->vega) (6.4.0)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\deeks\anaco
nda3\lib\site-packages (from jupyter-client>=5.3.4->notebook) (2.8.2)
Requirement already satisfied: pywin32>=1.0 in c:\users\deeks\anaconda3\lib
\site-packages (from jupyter-core>=4.6.1->notebook) (228)
Requirement already satisfied: pytz>=2017.3 in c:\users\deeks\anaconda3\lib
\site-packages (from pandas>=0.18->altair) (2021.3)
Requirement already satisfied: pywinpty>=0.5 in c:\users\deeks\anaconda3\li
b\site-packages (from terminado>=0.8.3->notebook) (0.5.7)
Requirement already satisfied: cffi>=1.0.0 in c:\users\deeks\anaconda3\lib
\site-packages (from argon2-cffi->notebook) (1.14.6)
Requirement already satisfied: pycparser in c:\users\deeks\anaconda3\lib\si
te-packages (from cffi>=1.0.0->argon2-cffi->notebook) (2.20)
Requirement already satisfied: ipython<8.0,>=7.23.1 in c:\users\deeks\anaco
nda3\lib\site-packages (from ipykernel->notebook) (7.29.0)
Requirement already satisfied: debugpy<2.0,>=1.0.0 in c:\users\deeks\anacon
da3\lib\site-packages (from ipykernel->notebook) (1.4.1)
Requirement already satisfied: matplotlib-inline<0.2.0,>=0.1.0 in c:\users
\deeks\anaconda3\lib\site-packages (from ipykernel->notebook) (0.1.2)
Requirement already satisfied: pickleshare in c:\users\deeks\anaconda3\lib
\site-packages (from ipython<8.0,>=7.23.1->ipykernel->notebook) (0.7.5)
Requirement already satisfied: colorama in c:\users\deeks\anaconda3\lib\sit
e-packages (from ipython<8.0,>=7.23.1->ipykernel->notebook) (0.4.4)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0
in c:\users\deeks\anaconda3\lib\site-packages (from ipython<8.0,>=7.23.1->i
pykernel->notebook) (3.0.20)
Requirement already satisfied: pygments in c:\users\deeks\anaconda3\lib\sit
e-packages (from ipython<8.0,>=7.23.1->ipykernel->notebook) (2.10.0)
Requirement already satisfied: jedi>=0.16 in c:\users\deeks\anaconda3\lib\s
ite-packages (from ipython<8.0,>=7.23.1->ipykernel->notebook) (0.18.0)
Requirement already satisfied: decorator in c:\users\deeks\anaconda3\lib\si
te-packages (from ipython<8.0,>=7.23.1->ipykernel->notebook) (5.1.0)
Requirement already satisfied: backcall in c:\users\deeks\anaconda3\lib\sit
e-packages (from ipython<8.0,>=7.23.1->ipykernel->notebook) (0.2.0)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in c:\users\deeks\anacon
da3\lib\site-packages (from jedi>=0.16->ipython<8.0,>=7.23.1->ipykernel->no
tebook) (0.8.2)
Requirement already satisfied: wcwidth in c:\users\deeks\anaconda3\lib\site
-packages (from prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython<8.0,>
```

```
=7.23.1->ipykernel->notebook) (0.2.5)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in c:\users\deeks
\anaconda3\lib\site-packages (from ipywidgets->jupyter<2.0.0,>=1.0.0->vega)
(1.0.0)
Requirement already satisfied: widgetsnbextension~=3.5.0 in c:\users\deeks
\anaconda3\lib\site-packages (from ipywidgets->jupyter<2.0.0,>=1.0.0->vega)
(3.5.1)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\deeks\anaconda3
\lib\site-packages (from jinja2->altair) (1.1.1)
Requirement already satisfied: testpath in c:\users\deeks\anaconda3\lib\sit
e-packages (from nbconvert->notebook) (0.5.0)
Requirement already satisfied: jupyterlab-pygments in c:\users\deeks\anacon
da3\lib\site-packages (from nbconvert->notebook) (0.1.2)
Requirement already satisfied: mistune<2,>=0.8.1 in c:\users\deeks\anaconda
3\lib\site-packages (from nbconvert->notebook) (0.8.4)
Requirement already satisfied: pandocfilters>=1.4.1 in c:\users\deeks\anaco
nda3\lib\site-packages (from nbconvert->notebook) (1.4.3)
Requirement already satisfied: nbclient<0.6.0,>=0.5.0 in c:\users\deeks\ana
conda3\lib\site-packages (from nbconvert->notebook) (0.5.3)
Requirement already satisfied: bleach in c:\users\deeks\anaconda3\lib\site-
packages (from nbconvert->notebook) (4.0.0)
Requirement already satisfied: defusedxml in c:\users\deeks\anaconda3\lib\s
ite-packages (from nbconvert->notebook) (0.7.1)
Requirement already satisfied: nest-asyncio in c:\users\deeks\anaconda3\lib
\site-packages (from nbclient<0.6.0,>=0.5.0->nbconvert->notebook) (1.5.1)
Requirement already satisfied: async-generator in c:\users\deeks\anaconda3
\lib\site-packages (from nbclient<0.6.0,>=0.5.0->nbconvert->notebook) (1.1
0)
Requirement already satisfied: webencodings in c:\users\deeks\anaconda3\lib
\site-packages (from bleach->nbconvert->notebook) (0.5.1)
Requirement already satisfied: packaging in c:\users\deeks\anaconda3\lib\si
te-packages (from bleach->nbconvert->notebook) (21.0)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\deeks\anaconda3
\lib\site-packages (from packaging->bleach->nbconvert->notebook) (3.0.4)
Requirement already satisfied: qtpy in c:\users\deeks\anaconda3\lib\site-pa
ckages (from qtconsole->jupyter<2.0.0,>=1.0.0->vega) (1.10.0)
```

#00,01-Data Creation,02-Exploratory Data Analysis

In [2]:
```
df_path=""
df_path2 = df_path + 'NewsData.csv'
df = pd.read_csv(df_path2, sep=';')
df.head()
```

Out[2]:

	File_Name	Content	Category	Complete_Filename
0	001.txt	Ad sales boost Time Warner profit\r\n\r\nQuart...	business	001.txt-business
1	002.txt	Dollar gains on Greenspan speech\r\n\r\nThe do...	business	002.txt-business
2	003.txt	Yukos unit buyer faces loan claim\r\n\r\nThe o...	business	003.txt-business
3	004.txt	High fuel prices hit BA's profits\r\n\r\nBriti...	business	004.txt-business
4	005.txt	Pernod takeover talk lifts Domecq\r\n\r\nShare...	business	005.txt-business

##Number of articles in each category

In [3]:

```
bars = alt.Chart(df).mark_bar(size=50).encode(
    x=alt.X("Category"),
    y=alt.Y("count():Q", axis=alt.Axis(title='Number of articles')),
    tooltip=[alt.Tooltip('count()', title='Number of articles'), 'Category'],
    color='Category'

)

text = bars.mark_text(
    align='center',
    baseline='bottom',
).encode(
    text='count()'
)

(bars + text).interactive().properties(
    height=300,
    width=700,
    title = "Number of articles in each category",
)
```

Out[3]:

```
##% of articles in each category
```

In [4]:

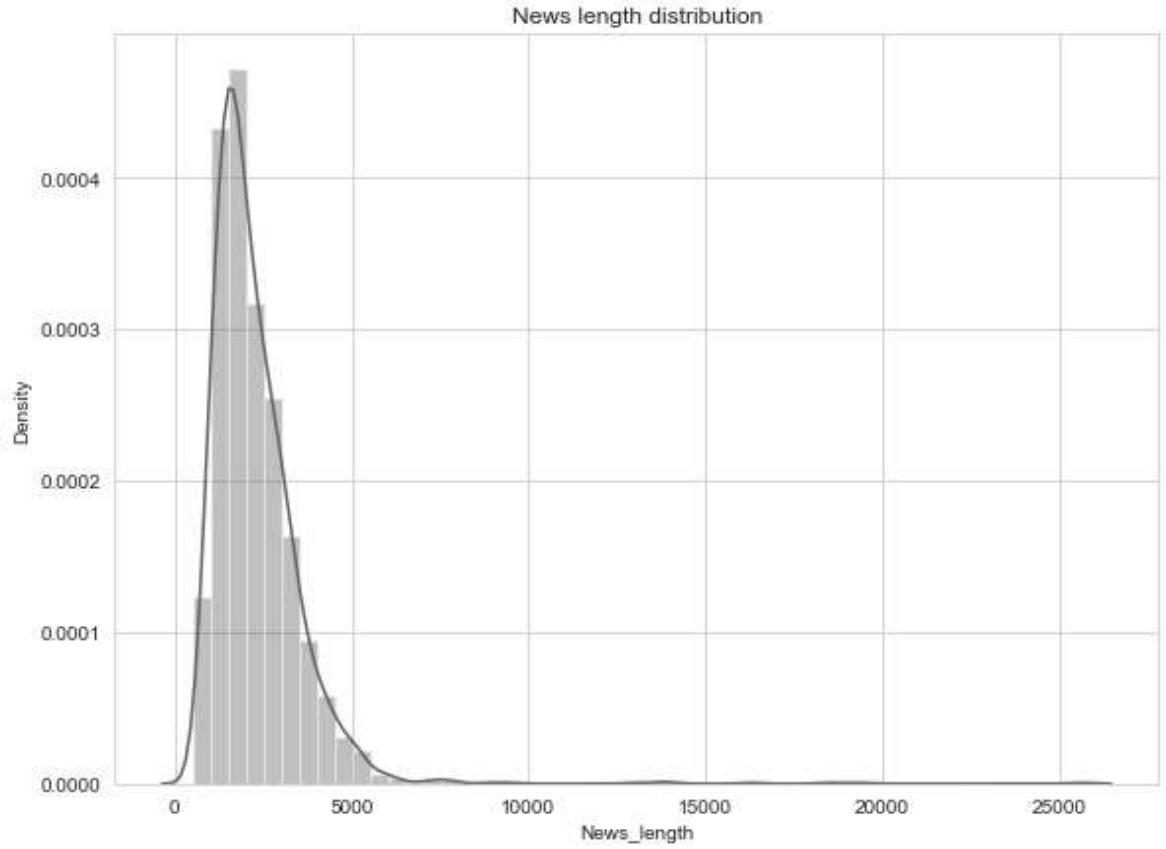
```
df['id'] = 1
df2 = pd.DataFrame(df.groupby('Category').count()['id']).reset_index()

bars = alt.Chart(df2).mark_bar(size=50).encode(
    x=alt.X('Category'),
    y=alt.Y('PercentOfTotal:Q', axis=alt.Axis(format='.0%', title='% of Artic
    color='Category'
).transform_window(
    TotalArticles='sum(id)',
    frame=[None, None]
).transform_calculate(
    PercentOfTotal="datum.id / datum.TotalArticles"
)

text = bars.mark_text(
    align='center',
    baseline='bottom',
    #dx=5  # Nudges text to right so it doesn't appear on top of the bar
).encode(
    text=alt.Text('PercentOfTotal:Q', format='.1%')
)

(bars + text).interactive().properties(
    height=300,
    width=700,
    title = "% of articles in each category",
)
```

Out[4]:

##News length by category

In [5]:

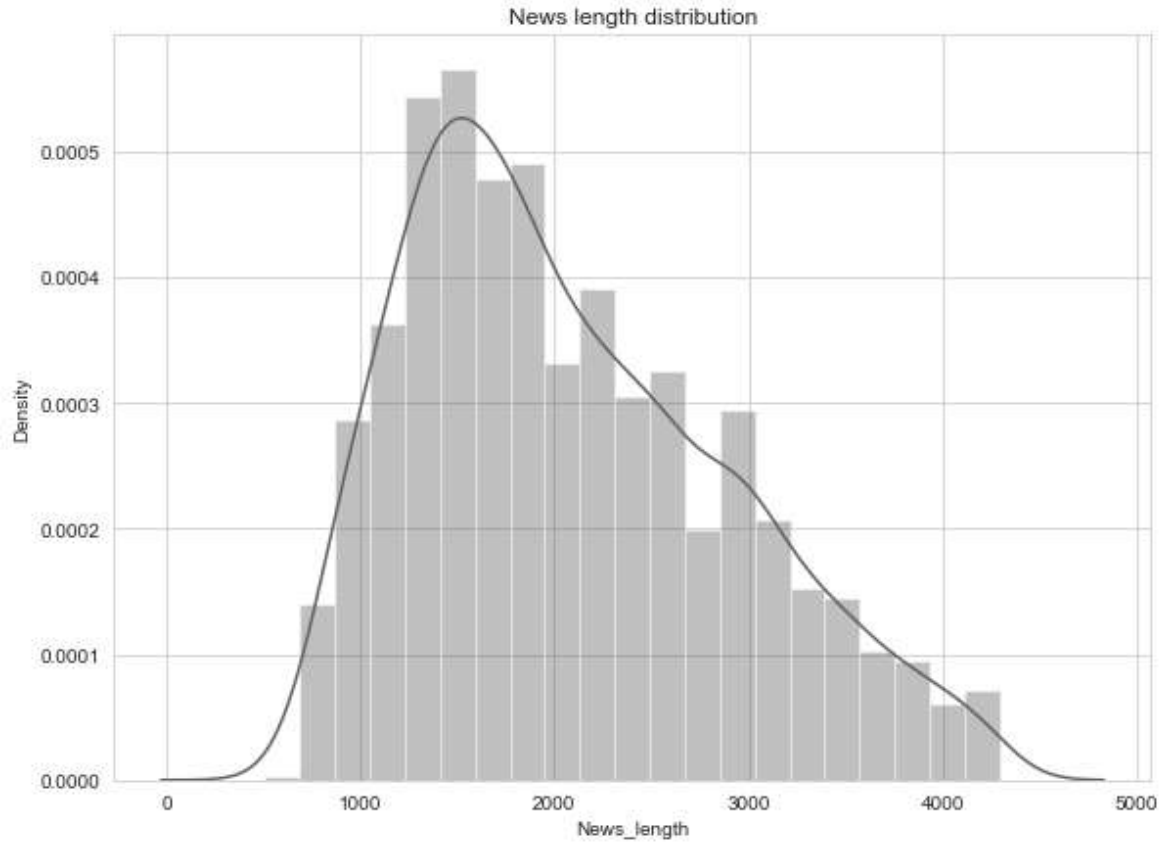
```
df['News_length'] = df['Content'].str.len()

plt.figure(figsize=(9.5,7))
sns.distplot(df['News_length']).set_title('News length distribution');
```



In [6]:
```
df['News_length'].describe()
```

Out[6]:
```
count     2225.000000
mean      2274.363596
std       1370.782663
min        506.000000
25%       1454.000000
50%       1978.000000
75%       2814.000000
max      25596.000000
Name: News_length, dtype: float64
```

In [7]:
```
quantile_95 = df['News_length'].quantile(0.95)
df_95 = df[df['News_length'] < quantile_95]
plt.figure(figsize=(9.5,7))
sns.distplot(df_95['News_length']).set_title('News length distribution');
```



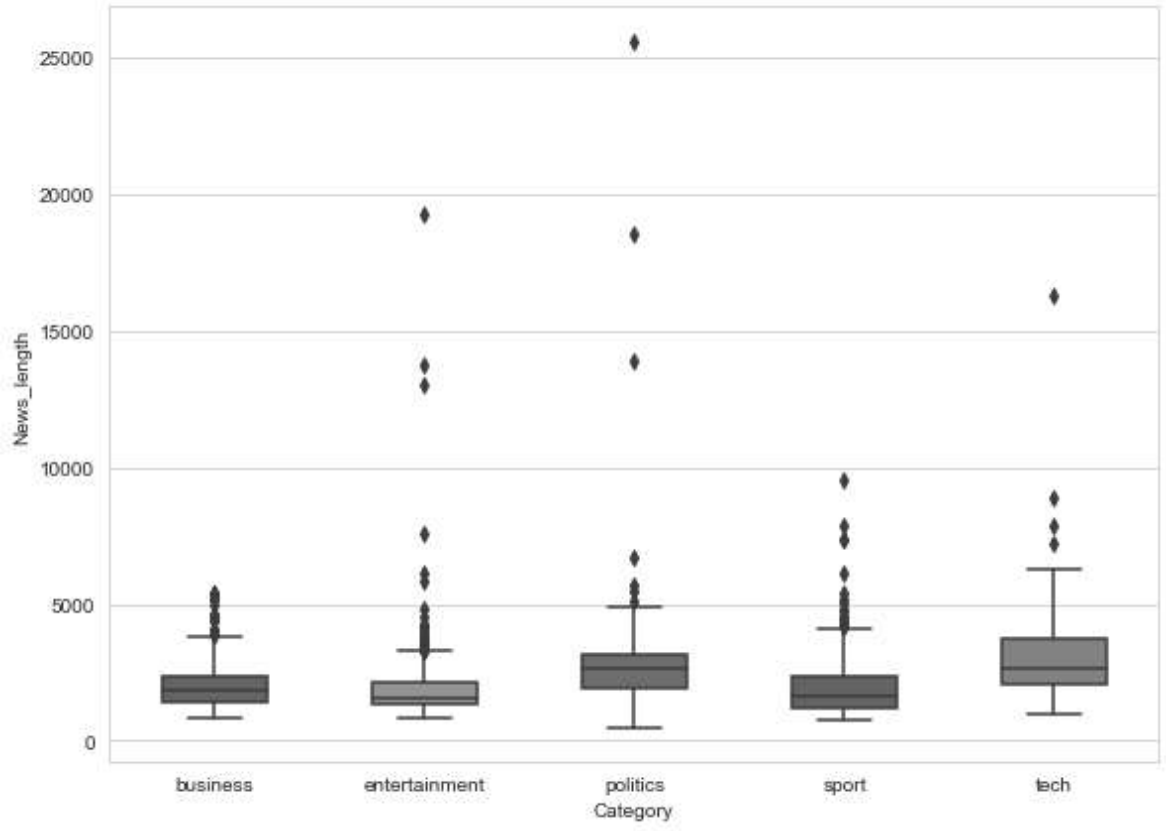
In [8]:
```
df_more10k = df[df['News_length'] > 10000]
len(df_more10k)
```

Out[8]:
```
7
```

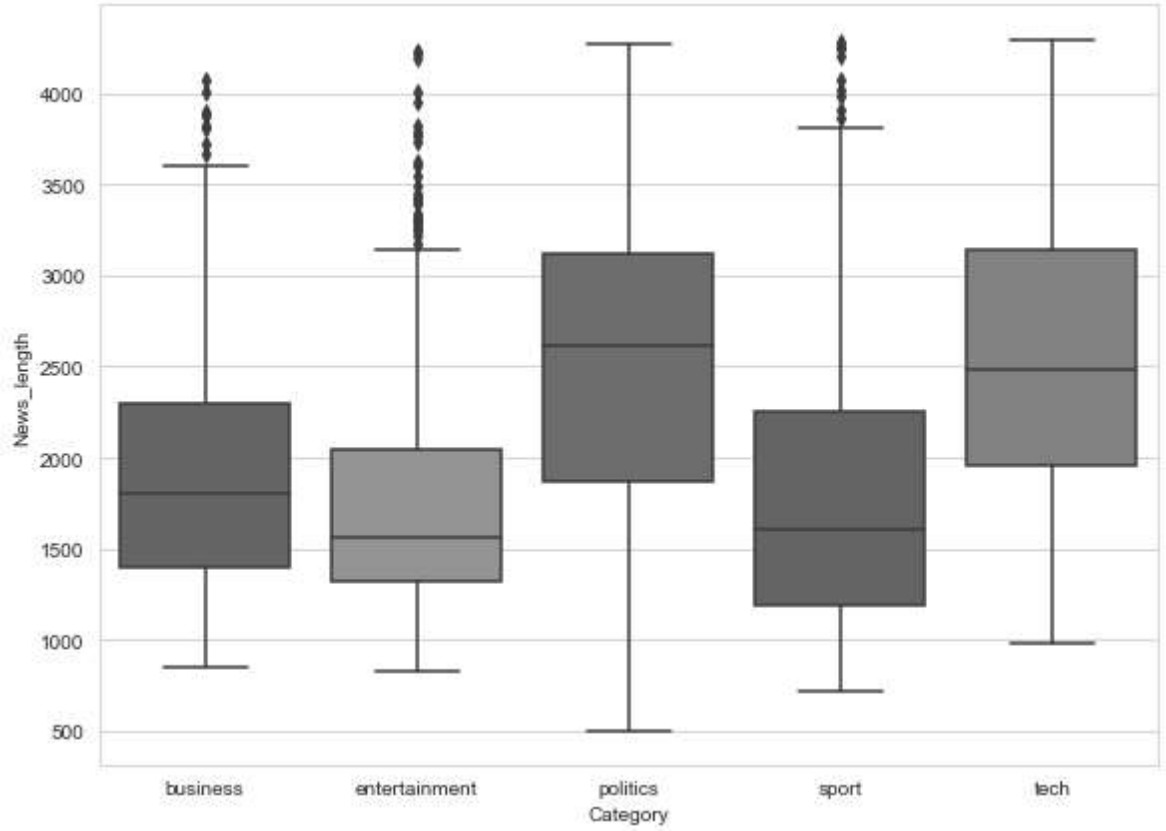
In [9]:
```
df_more10k['Content'].iloc[0]
```

```
is in trouble. Sure Robbie is talented and produces excellent material,
but this is not the best record.\r\n\r\nA sad day for music\r\n\r\nIt\'s
not the type of music I normally like, but even as a diehard rock fan, I
recognise that it is a good song and appelas to most people. That\'s why
it has been voted best song of the last 25 years. It\'s a good all-round
er. Just like Robbie.\r\n\r\nBest song in 25 years? Since 1980? I\'m con
fused. "Angels" isn\'t a bad song. It\'s a nice, catchy, formulaic anthe
m that ticks all the boxes. But this is not great music. If anything it
\'s regressive. Bland even. I suppose it\'s just more evidence of how re
dundant the Brit Awards have become.\r\n\r\nGranted angels is a good son
g, however it really wasn\'t up against any other proper competition. Th

e Queen\'s song was lackluster, and apart from Kate Bush, the other choi
ces were pathetic! Also, why weren\'t the Stones there, David Bowie, et
c, there are so many greater songs than Angels...I wonder if it was simp
ly the fact that Robbie wasn\'t getting more awards so they had to make
one up for him!\r\n\r\nBest song of the last 25 years? What a ridiculuou
s concept, and an even more ridiculous winner. Sigh. On the upside, at l
east it wasn\'t Bohemian Rhapsody, for which we should all be thankfu
l.\r\n\r\nAngels is without doubt a great song but I really don\'t think
```

In [10]:
```
plt.figure(figsize=(9.5,7))
sns.boxplot(data=df, x='Category', y='News_length', width=.5);
```



In [11]:
```
plt.figure(figsize=(9.5,7))
sns.boxplot(data=df_95, x='Category', y='News_length');
```



In [12]:
```
with open('NewsData.pickle', 'wb') as output:
    pickle.dump(df, output)
```

#03 - Feature Engineering

In [13]:
```
import pickle
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import chi2
import numpy as np
```

In [14]:

```
#path_df = "/home/lnc/0. Latest News Classifier/02. Exploratory Data Analysis
path_df = "NewsData.pickle"

with open(path_df, 'rb') as data:
    df = pickle.load(data)

df.head()
```

Out[14]:

	File_Name	Content	Category	Complete_Filename	id	News_length
0	001.txt	Ad sales boost Time Warner profit\r\n\r\nQuart...	business	001.txt-business	1	2569
1	002.txt	Dollar gains on Greenspan speech\r\n\r\nThe do...	business	002.txt-business	1	2257
2	003.txt	Yukos unit buyer faces loan claim\r\n\r\nThe o...	business	003.txt-business	1	1557
3	004.txt	High fuel prices hit BA's profits\r\n\r\nBriti...	business	004.txt-business	1	2421
4	005.txt	Pernod takeover talk lifts Domecq\r\n\r\nShare...	business	005.txt-business	1	1575

In [15]:
```
df.loc[1]['Content']
```

Out[15]:
```
'Dollar gains on Greenspan speech\r\n\r\nThe dollar has hit its highest lev
el against the euro in almost three months after the Federal Reserve head s
aid the US trade deficit is set to stabilise.\r\n\r\nAnd Alan Greenspan hig
hlighted the US government\'s willingness to curb spending and rising house
hold savings as factors which may help to reduce it. In late trading in New
York, the dollar reached $1.2871 against the euro, from $1.2974 on Thursda
y. Market concerns about the deficit has hit the greenback in recent month
s. On Friday, Federal Reserve chairman Mr Greenspan\'s speech in London ahe
ad of the meeting of G7 finance ministers sent the dollar higher after it h
ad earlier tumbled on the back of worse-than-expected US jobs data. "I thin
k the chairman\'s taking a much more sanguine view on the current account d
eficit than he\'s taken for some time," said Robert Sinche, head of currenc
y strategy at Bank of America in New York. "He\'s taking a longer-term vie
w, laying out a set of conditions under which the current account deficit c
an improve this year and next."\r\n\r\nWorries about the deficit concerns a
bout China do, however, remain. China\'s currency remains pegged to the dol
lar and the US currency\'s sharp falls in recent months have therefore made
Chinese export prices highly competitive. But calls for a shift in Beijing
\'s policy have fallen on deaf ears, despite recent comments in a major Chi
nese newspaper that the "time is ripe" for a loosening of the peg. The G7 m
eeting is thought unlikely to produce any meaningful movement in Chinese po
licy. In the meantime, the US Federal Reserve\'s decision on 2 February to
boost interest rates by a quarter of a point - the sixth such move in as ma
ny months - has opened up a differential with European rates. The half-poin
t window, some believe, could be enough to keep US assets looking more attr
active, and could help prop up the dollar. The recent falls have partly bee
n the result of big budget deficits, as well as the US\'s yawning current a
ccount gap, both of which need to be funded by the buying of US bonds and a
ssets by foreign firms and governments. The White House will announce its b
udget on Monday, and many commentators believe the deficit will remain at c
lose to half a trillion dollars.'
```

Text cleaning and preparation

Special character cleaning

In [16]:
```
# \r and \n
df['Content_Parsed_1'] = df['Content'].str.replace("\r", " ")
df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace("\n", " ")
df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace("    ", " ")

text = "Ms Preethika\'s"
print(text)

# " when quoting text
df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace('"', '')
```

```
Ms Preethika's
```

Upcase/downcase

In [17]:
```
# Lowercasing the text
df['Content_Parsed_2'] = df['Content_Parsed_1'].str.lower()
```

Punctuation signs

In [18]:
```
punctuation_signs = list("?:!.,;")
df['Content_Parsed_3'] = df['Content_Parsed_2']

for punct_sign in punctuation_signs:
    df['Content_Parsed_3'] = df['Content_Parsed_3'].str.replace(punct_sign, '
```

###Possessive pronouns

In [19]:
```
df['Content_Parsed_4'] = df['Content_Parsed_3'].str.replace("'s", "")
```

###Stemming and Lemmatization

In [20]:
```
# Download punkt and wordnet from NLTK
nltk.download('punkt')
print("------------------------------------------------------------")
nltk.download('wordnet')
```

```
------------------------------------------------------------

[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\deeks\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\deeks\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[20]:
```
True
```

In [21]:
```
# Lemmatizer is saved into an object
wordnet_lemmatizer = WordNetLemmatizer()

nltk.download('omw-1.4')

nrows = len(df)
lemmatized_text_list = []

for row in range(0, nrows):

    # Creating an empty list containing lemmatized words
    lemmatized_list = []

    # Saving the text and its words into an object
    text = df.loc[row]['Content_Parsed_4']
    text_words = text.split(" ")

    # Iterating through every word to lemmatize
    for word in text_words:
        lemmatized_list.append(wordnet_lemmatizer.lemmatize(word, pos="v"))

    # Joining the list
    lemmatized_text = " ".join(lemmatized_list)

    # Appending to the list containing the texts
    lemmatized_text_list.append(lemmatized_text)

df['Content_Parsed_5'] = lemmatized_text_list
```

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

###Stop words

In [22]:
```
# Download the stop words list
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\deeks\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[22]:
```
True
```

In [23]:
```
# Loading the stop words in english
stop_words = list(stopwords.words('english'))
stop_words[0:11]
```

Out[23]:
```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've"]
```

In [24]:
```
example = "myself watching movie"
word = "myself"

# The regular expression here:
regex = r"\b" + word + r"\b"  # we need to build it like that to work properl

re.sub(regex, "StopWord", example)
```

Out[24]:
```
'StopWord watching movie'
```

In [26]:
```
df['Content_Parsed_6'] = df['Content_Parsed_5']

for stop_word in stop_words:

    regex_stopword = r"\b" + stop_word + r"\b"
    df['Content_Parsed_6'] = df['Content_Parsed_6'].str.replace(regex_stopwor
```

Results of parsing

In [27]:
```
df.loc[5]['Content']
```

Out[27]:
```
'Japan narrowly escapes recession\r\n\r\nJapan\'s economy teetered on the b
rink of a technical recession in the three months to September, figures sho
w.\r\n\r\nRevised figures indicated growth of just 0.1% - and a similar-siz
ed contraction in the previous quarter. On an annual basis, the data sugges
ts annual growth of just 0.2%, suggesting a much more hesitant recovery tha
n had previously been thought. A common technical definition of a recession
is two successive quarters of negative growth.\r\n\r\nThe government was ke
en to play down the worrying implications of the data. "I maintain the view
that Japan\'s economy remains in a minor adjustment phase in an upward clim
b, and we will monitor developments carefully," said economy minister Heizo
Takenaka. But in the face of the strengthening yen making exports less comp
etitive and indications of weakening economic conditions ahead, observers w
ere less sanguine. "It\'s painting a picture of a recovery... much patchier
than previously thought," said Paul Sheard, economist at Lehman Brothers in
Tokyo. Improvements in the job market apparently have yet to feed through t
o domestic demand, with private consumption up just 0.2% in the third quart
er.'
```

In [28]:
```
df.loc[5]['Content_Parsed_1']
```

Out[28]:
```
"Japan narrowly escapes recession Japan's economy teetered on the brink of
a technical recession in the three months to September, figures show. Revis
ed figures indicated growth of just 0.1% - and a similar-sized contraction
in the previous quarter. On an annual basis, the data suggests annual growt
h of just 0.2%, suggesting a much more hesitant recovery than had previousl
y been thought. A common technical definition of a recession is two success
ive quarters of negative growth. The government was keen to play down the w
orrying implications of the data. I maintain the view that Japan's economy
remains in a minor adjustment phase in an upward climb, and we will monitor
developments carefully, said economy minister Heizo Takenaka. But in the fa
ce of the strengthening yen making exports less competitive and indications
of weakening economic conditions ahead, observers were less sanguine. It's
painting a picture of a recovery... much patchier than previously thought,
said Paul Sheard, economist at Lehman Brothers in Tokyo. Improvements in th
e job market apparently have yet to feed through to domestic demand, with p
rivate consumption up just 0.2% in the third quarter."
```

In [29]:
```
df.loc[5]['Content_Parsed_2']
```

Out[29]:
```
"japan narrowly escapes recession japan's economy teetered on the brink of
a technical recession in the three months to september, figures show. revis
ed figures indicated growth of just 0.1% - and a similar-sized contraction
in the previous quarter. on an annual basis, the data suggests annual growt
h of just 0.2%, suggesting a much more hesitant recovery than had previousl
y been thought. a common technical definition of a recession is two success
ive quarters of negative growth. the government was keen to play down the w
orrying implications of the data. i maintain the view that japan's economy
remains in a minor adjustment phase in an upward climb, and we will monitor
developments carefully, said economy minister heizo takenaka. but in the fa
ce of the strengthening yen making exports less competitive and indications
of weakening economic conditions ahead, observers were less sanguine. it's
painting a picture of a recovery... much patchier than previously thought,
said paul sheard, economist at lehman brothers in tokyo. improvements in th
e job market apparently have yet to feed through to domestic demand, with p
rivate consumption up just 0.2% in the third quarter."
```

In [30]:
```
df.loc[5]['Content_Parsed_3']
```

Out[30]:
```
"japan narrowly escapes recession japan's economy teetered on the brink of
a technical recession in the three months to september figures show revised
figures indicated growth of just 01% - and a similar-sized contraction in t
he previous quarter on an annual basis the data suggests annual growth of j
ust 02% suggesting a much more hesitant recovery than had previously been t
hought a common technical definition of a recession is two successive quart
ers of negative growth the government was keen to play down the worrying im
plications of the data i maintain the view that japan's economy remains in
a minor adjustment phase in an upward climb and we will monitor development
s carefully said economy minister heizo takenaka but in the face of the str
engthening yen making exports less competitive and indications of weakening
economic conditions ahead observers were less sanguine it's painting a pict
ure of a recovery much patchier than previously thought said paul sheard ec
onomist at lehman brothers in tokyo improvements in the job market apparent
ly have yet to feed through to domestic demand with private consumption up
just 02% in the third quarter"
```

In [31]:
```
df.loc[5]['Content_Parsed_4']
```

Out[31]:
```
'japan narrowly escapes recession japan economy teetered on the brink of a
technical recession in the three months to september figures show revised f
igures indicated growth of just 01% - and a similar-sized contraction in th
e previous quarter on an annual basis the data suggests annual growth of ju
st 02% suggesting a much more hesitant recovery than had previously been th
ought a common technical definition of a recession is two successive quarte
rs of negative growth the government was keen to play down the worrying imp
lications of the data i maintain the view that japan economy remains in a m
inor adjustment phase in an upward climb and we will monitor developments c
arefully said economy minister heizo takenaka but in the face of the streng
thening yen making exports less competitive and indications of weakening ec
onomic conditions ahead observers were less sanguine it painting a picture
of a recovery much patchier than previously thought said paul sheard econom
ist at lehman brothers in tokyo improvements in the job market apparently h
ave yet to feed through to domestic demand with private consumption up just
02% in the third quarter'
```

In [32]:
```
df.loc[5]['Content_Parsed_5']
```

Out[32]:
```
'japan narrowly escape recession japan economy teeter on the brink of a tec
hnical recession in the three months to september figure show revise figure
indicate growth of just 01% - and a similar-sized contraction in the previo
us quarter on an annual basis the data suggest annual growth of just 02% su
ggest a much more hesitant recovery than have previously be think a common
technical definition of a recession be two successive quarter of negative g
rowth the government be keen to play down the worry implications of the dat
a i maintain the view that japan economy remain in a minor adjustment phase
in an upward climb and we will monitor developments carefully say economy m
inister heizo takenaka but in the face of the strengthen yen make export le
ss competitive and indications of weaken economic condition ahead observers
be less sanguine it paint a picture of a recovery much patchier than previo
usly think say paul sheard economist at lehman brothers in tokyo improvemen
ts in the job market apparently have yet to fee through to domestic demand
with private consumption up just 02% in the third quarter'
```

In [33]:
```
df.loc[5]['Content_Parsed_6']
```

Out[33]:
```
'japan narrowly escape recession japan economy teeter   brink   technical r
ecession   three months  september figure show revise figure indicate growt
h   01% -   similar-sized contraction   previous quarter   annual basis  da
ta suggest annual growth   02% suggest  much  hesitant recovery   previousl
y  think  common technical definition   recession  two successive quarter
negative growth  government  keen  play   worry implications   data  mainta
in  view  japan economy remain   minor adjustment phase   upward climb    m
onitor developments carefully say economy minister heizo takenaka    face
strengthen yen make export less competitive  indications  weaken economic c
ondition ahead observers  less sanguine  paint  picture   recovery much pat
chier  previously think say paul sheard economist  lehman brothers  tokyo i
mprovements   job market apparently  yet  fee   domestic demand  private co
nsumption   02%   third quarter'
```

In [34]:
```
df.head(1)
```

Out[34]:

	File_Name	Content	Category	Complete_Filename	id	News_length	Content_Pars
0	001.txt	Ad sales boost Time Warner profit\r\n\r\nQuart...	business	001.txt-business	1	2569	Ad sales Time Warner Quarterl

In [35]:
```
list_columns = ["File_Name", "Category", "Complete_Filename", "Content", "Cor
df = df[list_columns]

df = df.rename(columns={'Content_Parsed_6': 'Content_Parsed'})

df.head()
```

Out[35]:

	File_Name	Category	Complete_Filename	Content	Content_Parsed
0	001.txt	business	001.txt-business	Ad sales boost Time Warner profit\r\n\r\nQuart...	ad sales boost time warner profit quarterly pr...
1	002.txt	business	002.txt-business	Dollar gains on Greenspan speech\r\n\r\nThe do...	dollar gain greenspan speech dollar hit hi...
2	003.txt	business	003.txt-business	Yukos unit buyer faces loan claim\r\n\r\nThe o...	yukos unit buyer face loan claim owners emba...
3	004.txt	business	004.txt-business	High fuel prices hit BA's profits\r\n\r\nBriti...	high fuel price hit ba profit british airways ...
4	005.txt	business	005.txt-business	Pernod takeover talk lifts Domecq\r\n\r\nShare...	pernod takeover talk lift domecq share uk dri...

Label Coding

In [36]:
```
category_codes = {
    'business': 0,
    'entertainment': 1,
    'politics': 2,
    'sport': 3,
    'tech': 4
}

# Category map
df['Category_Code'] = df['Category']
df = df.replace({'Category_Code':category_codes})

df.head()
```

Out[36]:

	File_Name	Category	Complete_Filename	Content	Content_Parsed	Category_C
0	001.txt	business	001.txt-business	Ad sales boost Time Warner profit\r\n\r\nQuart...	ad sales boost time warner profit quarterly pr...	
1	002.txt	business	002.txt-business	Dollar gains on Greenspan speech\r\n\r\nThe do...	dollar gain greenspan speech dollar hit hi...	
2	003.txt	business	003.txt-business	Yukos unit buyer faces loan claim\r\n\r\nThe o...	yukos unit buyer face loan claim owners emba...	
3	004.txt	business	004.txt-business	High fuel prices hit BA's profits\r\n\r\nBriti...	high fuel price hit ba profit british airways ...	
4	005.txt	business	005.txt-business	Pernod takeover talk lifts Domecq\r\n\r\nShare...	pernod takeover talk lift domecq share uk dri...	

Train - test split

In [37]:
```
X_train, X_test, y_train, y_test = train_test_split(df['Content_Parsed'],
                                                    df['Category_Code'],
                                                    test_size=0.15,
                                                    random_state=9)
```

##Text Representation

In [38]:
```
# Parameter election
ngram_range = (1,2)
min_df = 10
max_df = 1.
max_features = 300
```

In [39]:
```
tfidf = TfidfVectorizer(encoding='utf-8',
                        ngram_range=ngram_range,
                        stop_words=None,
                        lowercase=False,
                        max_df=max_df,
                        min_df=min_df,
                        max_features=max_features,
                        norm='l2',
                        sublinear_tf=True)

features_train = tfidf.fit_transform(X_train).toarray()
labels_train = y_train
print(features_train.shape)

features_test = tfidf.transform(X_test).toarray()
labels_test = y_test
print(features_test.shape)
```

```
(1891, 300)
(334, 300)
```

In [40]:

```
from sklearn.feature_selection import chi2
import numpy as np

for Product, category_id in sorted(category_codes.items()):
    features_chi2 = chi2(features_train, labels_train == category_id)
    indices = np.argsort(features_chi2[0])
    feature_names = np.array(tfidf.get_feature_names())[indices]
    unigrams = [v for v in feature_names if len(v.split(' ')) == 1]
    bigrams = [v for v in feature_names if len(v.split(' ')) == 2]
    print("# '{}' category:".format(Product))
    print("  . Most correlated unigrams:\n. {}".format('\n. '.join(unigrams[-
    print("  . Most correlated bigrams:\n. {}".format('\n. '.join(bigrams[-2:
    print("")
```

```
# 'business' category:
  . Most correlated unigrams:
. economic
. price
. economy
. growth
. bank
  . Most correlated bigrams:
. mr blair
. year old

# 'entertainment' category:
  . Most correlated unigrams:
. music
. tv
. award
. star
. film
  . Most correlated bigrams:
. mr blair
. prime minister

# 'politics' category:
  . Most correlated unigrams:
. minister
. party
. blair
. election
. labour
  . Most correlated bigrams:
. prime minister
. mr blair

# 'sport' category:
  . Most correlated unigrams:
. win
. season
. match
. champion
. cup
  . Most correlated bigrams:
. say mr
```

```
. year old

# 'tech' category:
  . Most correlated unigrams:
. digital
. technology
. computer
. software
. users
  . Most correlated bigrams:
. year old
. say mr
```

In [41]:
```
bigrams
```

Out[41]:
```
['tell bbc', 'last year', 'mr blair', 'prime minister', 'year old', 'say m
r']
```

In [42]:

```
# X_train
with open('Pickle/X_train.pickle', 'wb') as output:
    pickle.dump(X_train, output)

# X_test
with open('Pickle/X_test.pickle', 'wb') as output:
    pickle.dump(X_test, output)

# y_train
with open('Pickle/y_train.pickle', 'wb') as output:
    pickle.dump(y_train, output)

# y_test
with open('Pickle/y_test.pickle', 'wb') as output:
    pickle.dump(y_test, output)

# df
with open('Pickle/df.pickle', 'wb') as output:
    pickle.dump(df, output)

# features_train
with open('Pickle/features_train.pickle', 'wb') as output:
    pickle.dump(features_train, output)

# labels_train
with open('Pickle/labels_train.pickle', 'wb') as output:
    pickle.dump(labels_train, output)

# features_test
with open('Pickle/features_test.pickle', 'wb') as output:
    pickle.dump(features_test, output)

# labels_test
with open('Pickle/labels_test.pickle', 'wb') as output:
    pickle.dump(labels_test, output)

# TF-IDF object
with open('Pickle/tfidf.pickle', 'wb') as output:
    pickle.dump(tfidf, output)
```

04. Model Training

In [43]:
```
import pickle
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from pprint import pprint
from sklearn.model_selection import RandomizedSearchCV
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report, confusion_matrix, accuracy
from sklearn.model_selection import ShuffleSplit
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
```

In [44]:
```
# Dataframe
path_df = "Pickle/df.pickle"
with open(path_df, 'rb') as data:
    df = pickle.load(data)

# features_train
path_features_train = "Pickle/features_train.pickle"
with open(path_features_train, 'rb') as data:
    features_train = pickle.load(data)

# labels_train
path_labels_train = "Pickle/labels_train.pickle"
with open(path_labels_train, 'rb') as data:
    labels_train = pickle.load(data)

# features_test
path_features_test = "Pickle/features_test.pickle"
with open(path_features_test, 'rb') as data:
    features_test = pickle.load(data)

# labels_test
path_labels_test = "Pickle/labels_test.pickle"
with open(path_labels_test, 'rb') as data:
    labels_test = pickle.load(data)

print(features_train.shape)
print(features_test.shape)
```

```
(1891, 300)
(334, 300)
```

Random Forest

In [45]:

```
rf_0 = RandomForestClassifier(random_state = 8)

print('Parameters currently in use:\n')
pprint(rf_0.get_params())
```

```
Parameters currently in use:

{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 8,
 'verbose': 0,
 'warm_start': False}
```

In [46]:

```
# n_estimate
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 1000, num = 5

# max_feature
max_features = ['auto', 'sqrt']

#to find max_depth
max_depth = [int(x) for x in np.linspace(20, 100, num = 5)]
max_depth.append(None)

# min_samples_split
min_samples_split = [2, 5, 10]

# min_samples_leaf
min_samples_leaf = [1, 2, 4]

# bootstrap
bootstrap = [True, False]

# to Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}

pprint(random_grid)
```

```
{'bootstrap': [True, False],
 'max_depth': [20, 40, 60, 80, 100, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [200, 400, 600, 800, 1000]}
```

In [47]:
```
# create the base model to tune
rfc = RandomForestClassifier(random_state=8)

# random search
random_search = RandomizedSearchCV(estimator=rfc,
                                   param_distributions=random_grid,
                                   n_iter=50,
                                   scoring='accuracy',
                                   cv=3,
                                   verbose=1,
                                   random_state=8)

# Fit the random search model
random_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 50 candidates, totalling 150 fits
```

Out[47]:
```
RandomizedSearchCV(cv=3, estimator=RandomForestClassifier(random_state=8),
                   n_iter=50,
                   param_distributions={'bootstrap': [True, False],
                                        'max_depth': [20, 40, 60, 80, 100,
                                                      None],
                                        'max_features': ['auto', 'sqrt'],
                                        'min_samples_leaf': [1, 2, 4],
                                        'min_samples_split': [2, 5, 10],
                                        'n_estimators': [200, 400, 600, 80
0,
                                                         1000]},
                   random_state=8, scoring='accuracy', verbose=1)
```

In [48]:
```
print("The best hypepar from Random Search:")
print(random_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepar is:")
print(random_search.best_score_)
```

```
The best hypepar from Random Search:
{'n_estimators': 1000, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_
features': 'auto', 'max_depth': 80, 'bootstrap': True}

The mean accuracy of a model with these hypepar is:
0.9465868404061748
```

In [49]:
```
# parameter grid creation based on random search results
bootstrap = [False]
max_depth = [30, 40, 50]
max_features = ['sqrt']
min_samples_leaf = [1, 2, 4]
min_samples_split = [5, 10, 15]
n_estimators = [800]

param_grid = {
    'bootstrap': bootstrap,
    'max_depth': max_depth,
    'max_features': max_features,
    'min_samples_leaf': min_samples_leaf,
    'min_samples_split': min_samples_split,
    'n_estimators': n_estimators
}

# Create a base model
rfc = RandomForestClassifier(random_state=8)

# Manually create the splits in CV in order to be able to fix a random_state
cv_sets = ShuffleSplit(n_splits = 3, test_size = .33, random_state = 8)

# Instantiate the grid search model
grid_search = GridSearchCV(estimator=rfc,
                           param_grid=param_grid,
                           scoring='accuracy',
                           cv=cv_sets,
                           verbose=1)

# Fit the grid search to the data
grid_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 27 candidates, totalling 81 fits
```

Out[49]:
```
GridSearchCV(cv=ShuffleSplit(n_splits=3, random_state=8, test_size=0.33, tr
ain_size=None),
             estimator=RandomForestClassifier(random_state=8),
             param_grid={'bootstrap': [False], 'max_depth': [30, 40, 50],
                         'max_features': ['sqrt'],
                         'min_samples_leaf': [1, 2, 4],
                         'min_samples_split': [5, 10, 15],
                         'n_estimators': [800]},
             scoring='accuracy', verbose=1)
```

In [ ]:
```
print("The best hypepara from Grid Search are:")
print(grid_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(grid_search.best_score_)
```

In [ ]:
```
best_rfc = grid_search.best_estimator_
best_rfc
```

In [ ]:
```
best_rfc.fit(features_train, labels_train)
```

In [ ]:
```
rfc_pred = best_rfc.predict(features_test)
```

In [ ]:
```
# Train accuracy
print("The training accuracy is: ")
print(accuracy_score(labels_train, best_rfc.predict(features_train)))
```

In [ ]:
```
# Test accuracy
print("The test accuracy is: ")
print(accuracy_score(labels_test, rfc_pred))
```

In [ ]:
```
# Classification report
print("Classification report")
print(classification_report(labels_test,rfc_pred))
```

In [ ]:
```
aux_df = df[['Category', 'Category_Code']].drop_duplicates().sort_values('Cat
conf_matrix = confusion_matrix(labels_test, rfc_pred)
plt.figure(figsize=(12.8,6))
sns.heatmap(conf_matrix,
            annot=True,
            xticklabels=aux_df['Category'].values,
            yticklabels=aux_df['Category'].values,
            cmap="Blues")
plt.ylabel('Predicted')
plt.xlabel('Actual')
plt.title('Confusion matrix')
plt.show()
```

In [373]:
```
base_model = RandomForestClassifier(random_state = 8)
base_model.fit(features_train, labels_train)
accuracy_score(labels_test, base_model.predict(features_test))
```

Out[373]:
```
0.9281437125748503
```

In [374]:
```
best_rfc.fit(features_train, labels_train)
accuracy_score(labels_test, best_rfc.predict(features_test))
```

Out[374]:
```
0.9281437125748503
```

In [375]:
```
d = {
     'Model': 'Random Forest',
     'Training Set Accuracy': accuracy_score(labels_train, best_rfc.predict(f
     'Test Set Accuracy': accuracy_score(labels_test, rfc_pred)
}

df_models_rfc = pd.DataFrame(d, index=[0])
```

In [376]:
```
df_models_rfc
```

Out[376]:

	Model	Training Set Accuracy	Test Set Accuracy
0	Random Forest	1.0	0.928144

##Support Vector Machine

In [377]:
```
from sklearn import svm
svc_0 =svm.SVC(random_state=8)

print('Parameters currently in use:\n')
pprint(svc_0.get_params())
```

```
Parameters currently in use:

{'C': 1.0,
 'break_ties': False,
 'cache_size': 200,
 'class_weight': None,
 'coef0': 0.0,
 'decision_function_shape': 'ovr',
 'degree': 3,
 'gamma': 'scale',
 'kernel': 'rbf',
 'max_iter': -1,
 'probability': False,
 'random_state': 8,
 'shrinking': True,
 'tol': 0.001,
 'verbose': False}
```

In [378]:

```
# C
C = [.0001, .001, .01]

# gamma
gamma = [.0001, .001, .01, .1, 1, 10, 100]

# degree
degree = [1, 2, 3, 4, 5]

# kernel
kernel = ['linear', 'rbf', 'poly']

# probability
probability = [True]

# Create the random grid
random_grid = {'C': C,
               'kernel': kernel,
               'gamma': gamma,
               'degree': degree,
               'probability': probability
              }

pprint(random_grid)
```

```
{'C': [0.0001, 0.001, 0.01],
 'degree': [1, 2, 3, 4, 5],
 'gamma': [0.0001, 0.001, 0.01, 0.1, 1, 10, 100],
 'kernel': ['linear', 'rbf', 'poly'],
 'probability': [True]}
```

In [379]:
```
# First create the base model to tune
svc = svm.SVC(random_state=8)

# Definition of the random search
random_search = RandomizedSearchCV(estimator=svc,
                                   param_distributions=random_grid,
                                   n_iter=50,
                                   scoring='accuracy',
                                   cv=3,
                                   verbose=1,
                                   random_state=8)

# Fit the random search model
random_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 50 candidates, totalling 150 fits
```

Out[379]:
```
RandomizedSearchCV(cv=3, estimator=SVC(random_state=8), n_iter=50,
                   param_distributions={'C': [0.0001, 0.001, 0.01],
                                        'degree': [1, 2, 3, 4, 5],
                                        'gamma': [0.0001, 0.001, 0.01, 0.1,
1,
                                                  10, 100],
                                        'kernel': ['linear', 'rbf', 'pol
y'],
                                        'probability': [True]},
                   random_state=8, scoring='accuracy', verbose=1)
```

In [380]:
```
print("The best hypepara from Random Search are:")
print(random_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(random_search.best_score_)
```

```
The best hyperparameters from Random Search are:
{'probability': True, 'kernel': 'poly', 'gamma': 10, 'degree': 4, 'C': 0.0
1}

The mean accuracy of a model with these hyperparameters is:
0.9217358857612424
```

In [50]:

```
# Create the parameter grid based on the results of random search
C = [.0001, .001, .01, .1]
degree = [3, 4, 5]
gamma = [1, 10, 100]
probability = [True]

param_grid = [
  {'C': C, 'kernel':['linear'], 'probability':probability},
  {'C': C, 'kernel':['poly'], 'degree':degree, 'probability':probability},
  {'C': C, 'kernel':['rbf'], 'gamma':gamma, 'probability':probability}
]

# Create a base model
svc = svm.SVC(random_state=8)

# Manually create the splits in CV in order to be able to fix a random_state
cv_sets = ShuffleSplit(n_splits = 3, test_size = .33, random_state = 8)

# Instantiate the grid search model
grid_search = GridSearchCV(estimator=svc,
                           param_grid=param_grid,
                           scoring='accuracy',
                           cv=cv_sets,
                           verbose=1)

# Fit the grid search to the data
grid_search.fit(features_train, labels_train)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_20752/4129376872.py in <module>
     12
     13 # Create a base model
---> 14 svc = svm.SVC(random_state=8)
     15
     16 # Manually create the splits in CV in order to be able to fix a ran
dom_state (GridSearchCV doesn't have that argument)

NameError: name 'svm' is not defined
```

In [382]:

```
print("The best hypepara from Grid Search are:")
print(grid_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(grid_search.best_score_)
```

```
The best hyperparameters from Grid Search are:
{'C': 0.1, 'kernel': 'linear', 'probability': True}

The mean accuracy of a model with these hyperparameters is:
0.9498666666666665
```

In [383]:
```
best_svc = grid_search.best_estimator_
best_svc
```

Out[383]:
```
SVC(C=0.1, kernel='linear', probability=True, random_state=8)
```

In [384]:
```
best_svc.fit(features_train, labels_train)
```

Out[384]:
```
SVC(C=0.1, kernel='linear', probability=True, random_state=8)
```

In [385]:
```
svc_pred = best_svc.predict(features_test)
```

In [386]:
```
# Train accuracy
print("The training accuracy is: ")
print(accuracy_score(labels_train, best_svc.predict(features_train)))
```

```
The training accuracy is:
0.9592808038075092
```

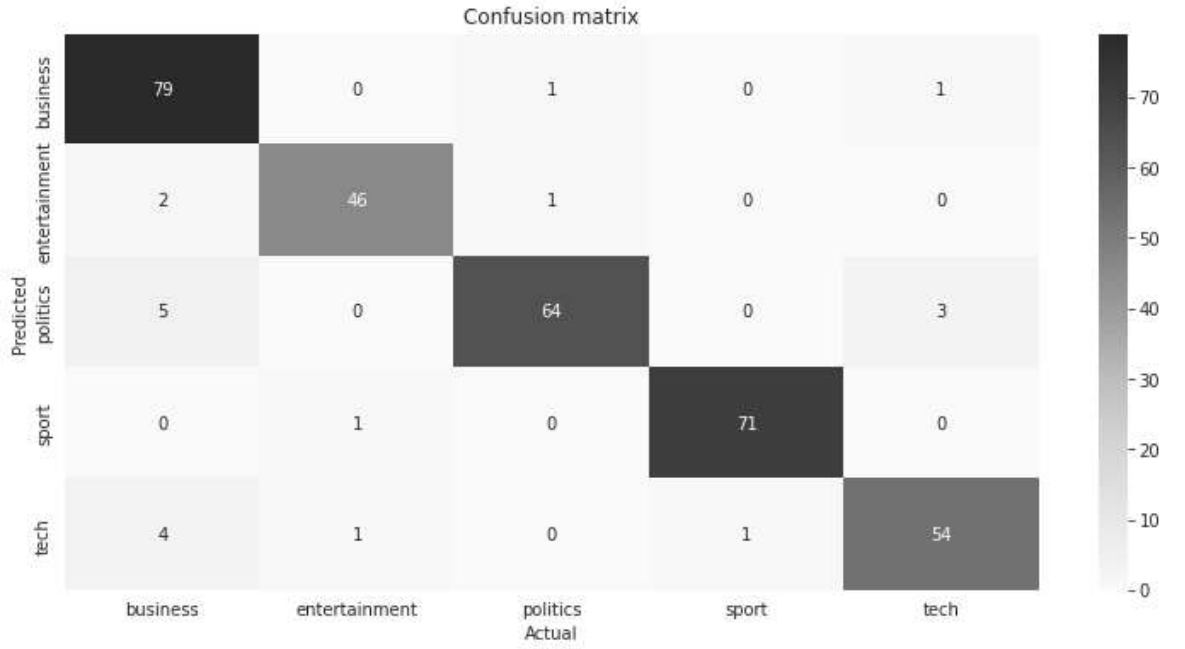
In [387]:
```
# Test accuracy
print("The test accuracy is: ")
print(accuracy_score(labels_test, svc_pred))
```

```
The test accuracy is:
0.9401197604790419
```

In [388]:
```
# Classification report
print("Classification report")
print(classification_report(labels_test,svc_pred))
```

```
Classification report
              precision    recall  f1-score   support

           0       0.88      0.98      0.92        81
           1       0.96      0.94      0.95        49
           2       0.97      0.89      0.93        72
           3       0.99      0.99      0.99        72
           4       0.93      0.90      0.92        60

    accuracy                           0.94       334
   macro avg       0.94      0.94      0.94       334
weighted avg       0.94      0.94      0.94       334
```

In [389]:
```
aux_df = df[['Category', 'Category_Code']].drop_duplicates().sort_values('Cat
conf_matrix = confusion_matrix(labels_test, svc_pred)
plt.figure(figsize=(12.8,6))
sns.heatmap(conf_matrix,
            annot=True,
            xticklabels=aux_df['Category'].values,
            yticklabels=aux_df['Category'].values,
            cmap="Blues")
plt.ylabel('Predicted')
plt.xlabel('Actual')
plt.title('Confusion matrix')
plt.show()
```



In [390]:
```
base_model = svm.SVC(random_state = 8)
base_model.fit(features_train, labels_train)
accuracy_score(labels_test, base_model.predict(features_test))
```

Out[390]: 0.9550898203592815

In [391]:
```
best_svc.fit(features_train, labels_train)
accuracy_score(labels_test, best_svc.predict(features_test))
```

Out[391]: 0.9401197604790419

In [392]:
```
d = {
     'Model': 'SVM',
     'Training Set Accuracy': accuracy_score(labels_train, best_svc.predict(f
     'Test Set Accuracy': accuracy_score(labels_test, svc_pred)
}

df_models_svc = pd.DataFrame(d, index=[0])
df_models_svc
```

Out[392]:

	Model	Training Set Accuracy	Test Set Accuracy
0	SVM	0.959281	0.94012

In [393]:
```
with open('Models/best_svc.pickle', 'wb') as output:
    pickle.dump(best_svc, output)

with open('Models/df_models_svc.pickle', 'wb') as output:
    pickle.dump(df_models_svc, output)
```

##KNN

In [394]:
```
from sklearn.neighbors import KNeighborsClassifier
knnc_0 =KNeighborsClassifier()

print('Parameters currently in use:\n')
pprint(knnc_0.get_params())
```

```
Parameters currently in use:

{'algorithm': 'auto',
 'leaf_size': 30,
 'metric': 'minkowski',
 'metric_params': None,
 'n_jobs': None,
 'n_neighbors': 5,
 'p': 2,
 'weights': 'uniform'}
```

In [395]:
```
# Create parameter grid
n_neighbors = [int(x) for x in np.linspace(start = 1, stop = 500, num = 100)]

param_grid = {'n_neighbors': n_neighbors}

# Create base model
knnc = KNeighborsClassifier()

# Manually create the splits in CV in order to be able to fix a random_state
cv_sets = ShuffleSplit(n_splits = 3, test_size = .33, random_state = 8)

# Instantiate the grid search model
grid_search = GridSearchCV(estimator=knnc,
                           param_grid=param_grid,
                           scoring='accuracy',
                           cv=cv_sets,
                           verbose=1)

# Fit the grid search to the data
grid_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 100 candidates, totalling 300 fits
```

Out[395]:
```
GridSearchCV(cv=ShuffleSplit(n_splits=3, random_state=8, test_size=0.33, tr
ain_size=None),
             estimator=KNeighborsClassifier(),
             param_grid={'n_neighbors': [1, 6, 11, 16, 21, 26, 31, 36, 41,
46,
                                         51, 56, 61, 66, 71, 76, 81, 86, 9
1, 96,
                                         101, 106, 111, 116, 121, 127, 132,
137,
                                         142, 147, ...]},
             scoring='accuracy', verbose=1)
```

In [396]:
```
print("The best hypepara from Grid Search are:")
print(grid_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(grid_search.best_score_)
```

```
The best hyperparameters from Grid Search are:
{'n_neighbors': 6}

The mean accuracy of a model with these hyperparameters is:
0.9477333333333333
```

In [397]:
```
n_neighbors = [1,2,3,4,5,6,7,8,9,10,11]
param_grid = {'n_neighbors': n_neighbors}

knnc = KNeighborsClassifier()
cv_sets = ShuffleSplit(n_splits = 3, test_size = .33, random_state = 8)

grid_search = GridSearchCV(estimator=knnc,
                           param_grid=param_grid,
                           scoring='accuracy',
                           cv=cv_sets,
                           verbose=1)

grid_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 11 candidates, totalling 33 fits
```

Out[397]:
```
GridSearchCV(cv=ShuffleSplit(n_splits=3, random_state=8, test_size=0.33, tr
ain_size=None),
             estimator=KNeighborsClassifier(),
             param_grid={'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1
1]},
             scoring='accuracy', verbose=1)
```

In [398]:
```
print("The best hypepara from Grid Search are:")
print(grid_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(grid_search.best_score_)
```

```
The best hyperparameters from Grid Search are:
{'n_neighbors': 6}

The mean accuracy of a model with these hyperparameters is:
0.9477333333333333
```

In [399]:
```
best_knnc = grid_search.best_estimator_
best_knnc
```

Out[399]:
```
KNeighborsClassifier(n_neighbors=6)
```

In [400]:
```
best_knnc.fit(features_train, labels_train)
```

Out[400]:
```
KNeighborsClassifier(n_neighbors=6)
```

In [401]:
```
knnc_pred = best_knnc.predict(features_test)
```

In [402]:
```
# Train accuracy
print("The training accuracy is: ")
print(accuracy_score(labels_train, best_knnc.predict(features_train)))
```

```
The training accuracy is:
0.9598096245372819
```

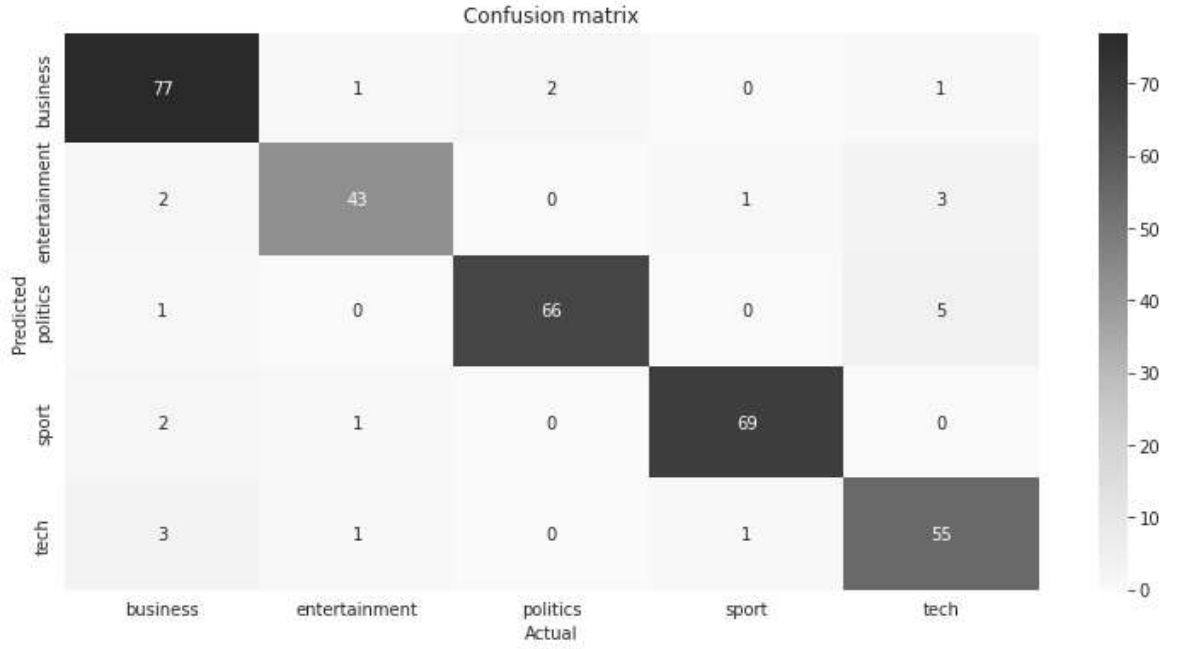
In [403]:
```
# Test accuracy
print("The test accuracy is: ")
print(accuracy_score(labels_test, knnc_pred))
```

```
The test accuracy is:
0.9281437125748503
```

In [404]:
```
# Classification report with precision recall f1-score and support
print("Classification report")
print(classification_report(labels_test,knnc_pred))
```

```
Classification report
              precision    recall  f1-score   support

           0       0.91      0.95      0.93        81
           1       0.93      0.88      0.91        49
           2       0.97      0.92      0.94        72
           3       0.97      0.96      0.97        72
           4       0.86      0.92      0.89        60

    accuracy                           0.93       334
   macro avg       0.93      0.92      0.93       334
weighted avg       0.93      0.93      0.93       334
```

In [405]:
```
aux_df = df[['Category', 'Category_Code']].drop_duplicates().sort_values('Cat
conf_matrix = confusion_matrix(labels_test, knnc_pred)
plt.figure(figsize=(12.8,6))
sns.heatmap(conf_matrix,
            annot=True,
            xticklabels=aux_df['Category'].values,
            yticklabels=aux_df['Category'].values,
            cmap="Blues")
plt.ylabel('Predicted')
plt.xlabel('Actual')
plt.title('Confusion matrix')
plt.show()
```



In [406]:
```
base_model = KNeighborsClassifier()
base_model.fit(features_train, labels_train)
accuracy_score(labels_test, base_model.predict(features_test))
```

Out[406]: 0.9341317365269461

In [407]:
```
best_knnc.fit(features_train, labels_train)
accuracy_score(labels_test, best_knnc.predict(features_test))
```

Out[407]: 0.9281437125748503

In [408]:
```
d = {
     'Model': 'KNN',
     'Training Set Accuracy': accuracy_score(labels_train, best_knnc.predict(
     'Test Set Accuracy': accuracy_score(labels_test, knnc_pred)
}

df_models_knnc = pd.DataFrame(d, index=[0])
```

In [409]:
```
df_models_knnc
```

Out[409]:

	Model	Training Set Accuracy	Test Set Accuracy
0	KNN	0.95981	0.928144

In [410]:
```
with open('Models/best_knnc.pickle', 'wb') as output:
    pickle.dump(best_knnc, output)

with open('Models/df_models_knnc.pickle', 'wb') as output:
    pickle.dump(df_models_knnc, output)
```

##Model Selection

In [411]:
```
path_pickles = "Models/"

list_pickles = [
    "df_models_knnc.pickle",
    "df_models_rfc.pickle",
    "df_models_svc.pickle"
]

df_summary = pd.DataFrame()

for pickle_ in list_pickles:

    path = path_pickles + pickle_

    with open(path, 'rb') as data:
        df = pickle.load(data)

    df_summary = df_summary.append(df)

df_summary = df_summary.reset_index().drop('index', axis=1)
df_summary
```

Out[411]:

	Model	Training Set Accuracy	Test Set Accuracy
0	KNN	0.959810	0.928144
1	Random Forest	1.000000	0.928144
2	SVM	0.959281	0.940120

In [412]:
```
df_summary.sort_values('Test Set Accuracy', ascending=False)
```

Out[412]:

	Model	Training Set Accuracy	Test Set Accuracy
2	SVM	0.959281	0.940120
0	KNN	0.959810	0.928144
1	Random Forest	1.000000	0.928144

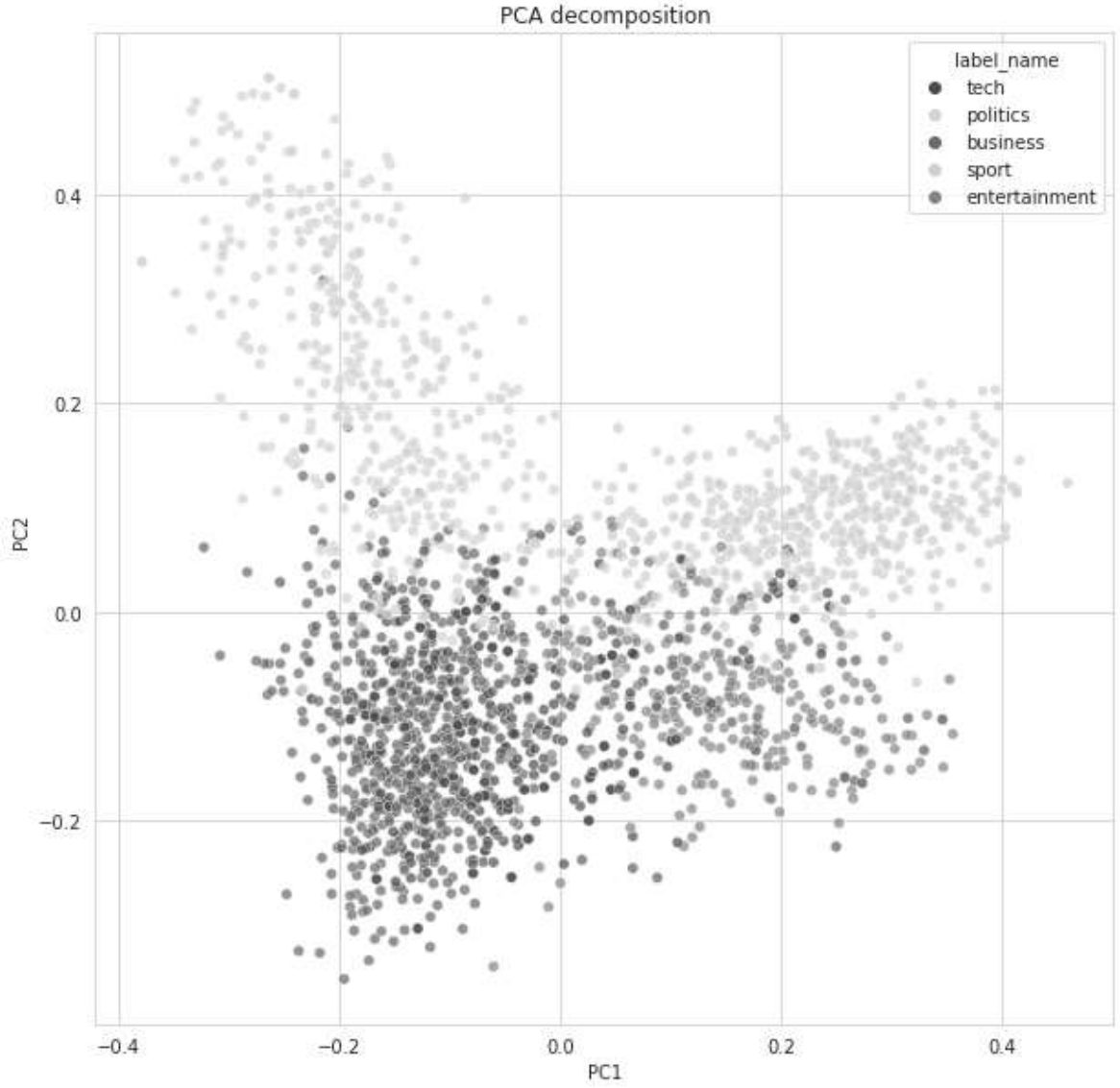
In [413]:
```
features = np.concatenate((features_train,features_test), axis=0)
labels = np.concatenate((labels_train,labels_test), axis=0)
print(features.shape)
print(labels.shape)
```

```
(2225, 300)
(2225,)
```

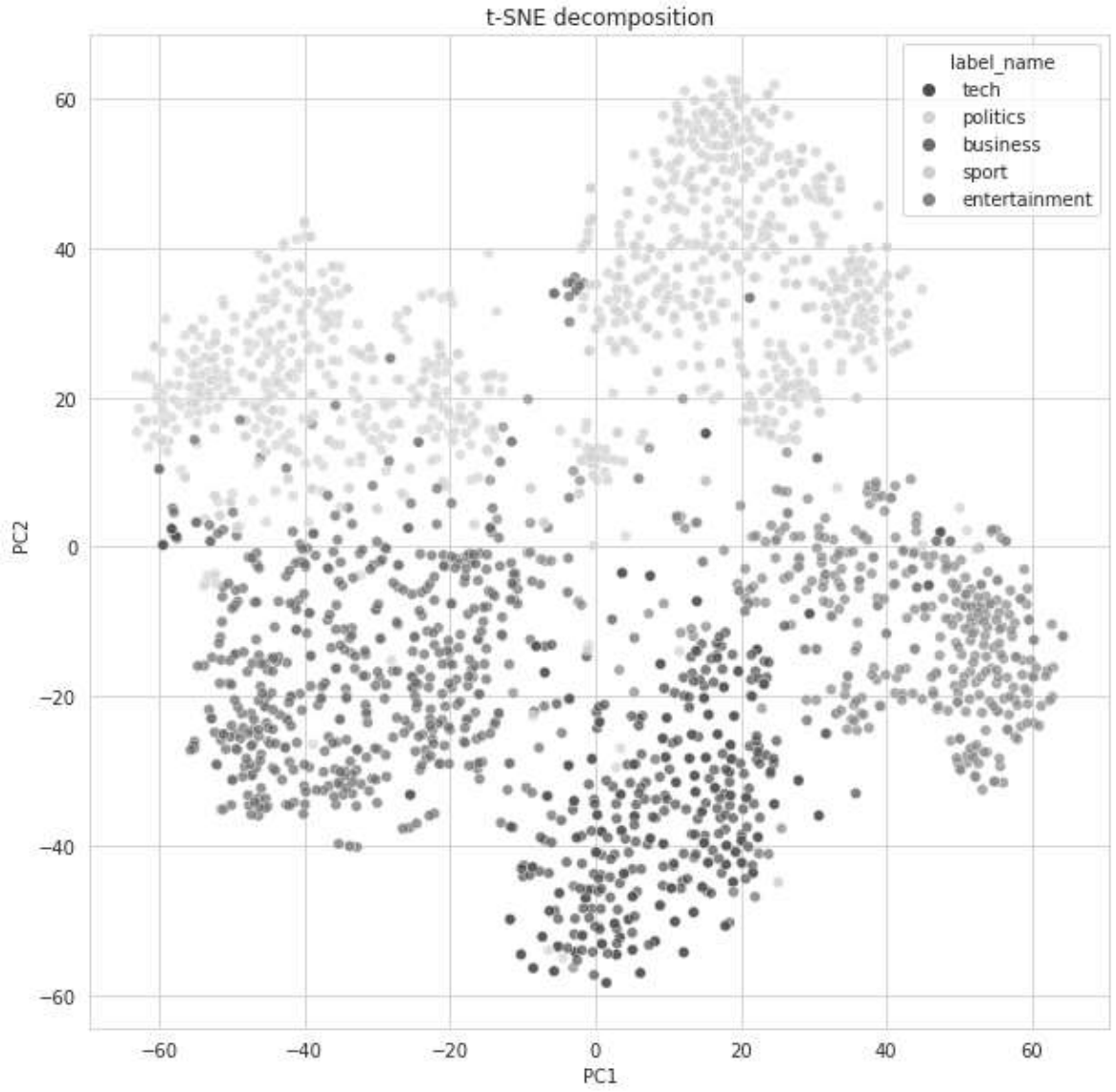
In [414]:

```
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE

def plot_dim_red(model, features, labels, n_components=2):

    # Creation of the model
    if (model == 'PCA'):
        mod = PCA(n_components=n_components)
        title = "PCA decomposition"  # for the plot

    elif (model == 'TSNE'):
        mod = TSNE(n_components=2)
        title = "t-SNE decomposition"

    else:
        return "Error"

    # Fit and transform the features
    principal_components = mod.fit_transform(features)

    # Put them into a dataframe
    df_features = pd.DataFrame(data=principal_components,
                     columns=['PC1', 'PC2'])

    # Now we have to paste each row's label and its meaning
    # Convert labels array to df
    df_labels = pd.DataFrame(data=labels,
                             columns=['label'])

    df_full = pd.concat([df_features, df_labels], axis=1)
    df_full['label'] = df_full['label'].astype(str)

    # Get labels name
    category_names = {
        "0": 'business',
        "1": 'entertainment',
        "2": 'politics',
        "3": 'sport',
        "4": 'tech'
    }

    # And map labels
    df_full['label_name'] = df_full['label']
    df_full = df_full.replace({'label_name':category_names})

    # Plot
    plt.figure(figsize=(10,10))
    sns.scatterplot(x='PC1',
                    y='PC2',
                    hue="label_name",
                    data=df_full,
                    palette=["red", "pink", "royalblue", "greenyellow", "ligh
                    alpha=.7).set_title(title);
```

In [415]:
```
plot_dim_red("PCA",
             features=features,
             labels=labels,
             n_components=2)
```



In [416]:
```
plot_dim_red("TSNE",
             features=features,
             labels=labels,
             n_components=2)
```



In [417]:

```
# Dataframe
path_df = "Pickle/df.pickle"
with open(path_df, 'rb') as data:
    df = pickle.load(data)

# SVM Model
path_model = "Models/best_svc.pickle"
with open(path_model, 'rb') as data:
    svc_model = pickle.load(data)

# Category mapping dictionary
category_codes = {
    'business': 0,
    'entertainment': 1,
    'politics': 2,
    'sport': 3,
    'tech': 4
}

category_names = {
    0: 'business',
    1: 'entertainment',
    2: 'politics',
    3: 'sport',
    4: 'tech'
}
```

In [418]:

```
predictions = svc_model.predict(features_test)
# Indexes of the test set
index_X_test = X_test.index

print(index_X_test)


# We get them from the original df
df_test = df.loc[index_X_test]

# Add the predictions
df_test['Prediction'] = predictions

# Clean columns
df_test = df_test[['Content', 'Category', 'Category_Code', 'Prediction']]

# Decode
df_test['Category_Predicted'] = df_test['Prediction']
df_test = df_test.replace({'Category_Predicted':category_names})

# Clean columns again
df_test = df_test[['Content', 'Category', 'Category_Predicted']]
df_test.head()
```

```
Int64Index([1691, 1103,  477,  197,  475,  162,  887,  307, 1336, 1679,
            ...
            1567, 2130, 1216, 1135,  359,  393, 1746,  444, 2215,  733],
           dtype='int64', length=334)
```

Out[418]:

	Content	Category	Category_Predicted
1691	Ireland call up uncapped Campbell\r\n\r\nUlste...	sport	sport
1103	Gurkhas to help tsunami victims\r\n\r\nBritain...	politics	business
477	Egypt and Israel seal trade deal\r\n\r\nIn a s...	business	business
197	Cairn shares up on new oil find\r\n\r\nShares ...	business	business
475	Saudi NCCI's shares soar\r\n\r\nShares in Saud...	business	business

In [419]:

```
condition = (df_test['Category'] != df_test['Category_Predicted'])

df_misclassified = df_test[condition]

df_misclassified.head(3)
```

Out[419]:

	Content	Category	Category_Predicted
1103	Gurkhas to help tsunami victims\r\n\r\nBritain...	politics	business
1880	Half-Life 2 sweeps Bafta awards\r\n\r\nPC firs...	tech	entertainment
2137	Junk e-mails on relentless rise\r\n\r\nSpam tr...	tech	business

In [420]:
```
def output_article(row_article):
    print('Actual Category: %s' %(row_article['Category']))
    print('Predicted Category: %s' %(row_article['Category_Predicted']))
    print('----------------------------------------')
    print('Text: ')
    print('%s' %(row_article['Content']))
```

In [421]:
```
import random

random.seed(8)
list_samples = random.sample(list(df_misclassified.index), 3)
list_samples
```

Out[421]: [956, 1339, 1205]

In [422]:
```
output_article(df_misclassified.loc[list_samples[0]])
```

```
Actual Category: politics
Predicted Category: tech
----------------------------------------
Text:
Assembly ballot papers 'missing'

Hundreds of ballot papers for the regional assembly referendum in the North
East have "disappeared".

Royal Mail says it is investigating the situation, which has meant about 30
0 homes in County Durham are not receiving voting packs. Officials at Darli
ngton Council are now in a race against time to try and rectify the situati
on. The all-postal votes of about two million electors are due to be handed
in by 4 November. A spokesman for Darlington Council said: "We have sent ou
t the ballot papers, the problem is with Royal Mail. "Somewhere along the l
ine, something has gone wrong and these ballot papers have not been deliver
ed. "The Royal Mail is investigating to see if they can find out what the p
roblem is."

A spokeswoman for Royal Mail said: "We are investigating a problem with the
delivery route in the Mowden area of Darlington. "This is affecting several
hundred properties, which have failed to receive ballot papers. "We are wor
king closely with the council and will do all we can to help rectify the pr
oblem. "No-one will not receive their ballot paper as special hand deliveri
es will take place where necessary. "We are unaware of any other problems o
f this kind to do with the regional assembly vote."

The Darlington Council spokesman added: "Initially we had complaints from a
couple of residents in Mowden to say they thought they should have had thei
r ballot papers by now. "We then made further investigations and it became
clear this was a bigger issue." A spokeswoman for the Electoral Commission
told BBC News Online that letters were being sent out to those homes affect
ed. She said the commission was satisfied that measures had been put in pla
ce to ensure all voters received ballot papers in time. So far a total of 5
69,072 ballot envelopes have been scanned by bar code at counting offices a
cross the North East.
```

In [423]:
```
output_article(df_misclassified.loc[list_samples[1]])
```

```
Actual Category: sport
Predicted Category: entertainment
-----------------------------------------
Text:
Holmes feted with further honour

Double Olympic champion Kelly Holmes has been voted European Athletics (EA
A) woman athlete of 2004 in the governing body's annual poll.

The Briton, made a dame in the New Year Honours List for taking 800m and 1,
500m gold, won vital votes from the public, press and EAA member federation
s. She is only the second British woman to land the title after- Sally Gunn
ell won for her world 400m hurdles win in 1993. Swedish triple jumper Chris
tian Olsson was voted male athlete of the year. The accolade is the latest
in a long list of awards that Holmes has received since her success in Athe
ns. In addition to becoming a dame, she was also named the BBC Sports Perso
nality of the Year in December. Her gutsy victory in the 800m also earned h
er the International Association of Athletics Federations' award for the be
st women's performance in the world for 2004. And she scooped two awards at
the British Athletics Writers' Association annual dinner in October.
```

In [424]:
```
output_article(df_misclassified.loc[list_samples[2]])
```

```
Actual Category: politics
Predicted Category: tech
-----------------------------------------
Text:
MPs issued with Blackberry threat

MPs will be thrown out of the Commons if they use Blackberries in the chamb
er Speaker Michael Martin has ruled.

The Â£200 handheld computers can be used as a phone, pager or to send e-mai
ls. The devices gained new prominence this week after Alastair Campbell use
d his to accidentally send an expletive-laden message to a Newsnight journa
list. Mr Martin revealed some MPs had been using their Blackberries during
debates and he also cautioned members against using hidden earpieces.

The use of electronic devices in the Commons chamber has long been frowned
on. The sound of a mobile phone or a pager can result in a strong rebuke fr
om either the Speaker or his deputies. The Speaker chairs debates in the Co
mmons and is charged with ensuring order in the chamber and enforcing rules
and conventions of the House. He or she is always an MP chosen by colleague
s who, once nominated, gives up all party political allegiances.
```

In [425]:

```
path_models = "Models/"

# SVM
path_svm = path_models + 'best_svc.pickle'
with open(path_svm, 'rb') as data:
    svc_model = pickle.load(data)

path_tfidf = "Pickle/tfidf.pickle"
with open(path_tfidf, 'rb') as data:
    tfidf = pickle.load(data)

category_codes = {
    'business': 0,
    'entertainment': 1,
    'politics': 2,
    'sport': 3,
    'tech': 4
}
```

In [426]:

```
punctuation_signs = list("?:!.,;")
stop_words = list(stopwords.words('english'))

def create_features_from_text(text):

    # Dataframe creation
    lemmatized_text_list = []
    df = pd.DataFrame(columns=['Content'])
    df.loc[0] = text
    df['Content_Parsed_1'] = df['Content'].str.replace("\r", " ")
    df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace("\n", " ")
    df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace("    ", " ")
    df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace('"', '')
    df['Content_Parsed_2'] = df['Content_Parsed_1'].str.lower()
    df['Content_Parsed_3'] = df['Content_Parsed_2']
    for punct_sign in punctuation_signs:
        df['Content_Parsed_3'] = df['Content_Parsed_3'].str.replace(punct_sig
    df['Content_Parsed_4'] = df['Content_Parsed_3'].str.replace("'s", "")
    wordnet_lemmatizer = WordNetLemmatizer()
    lemmatized_list = []
    text = df.loc[0]['Content_Parsed_4']
    text_words = text.split(" ")
    for word in text_words:
        lemmatized_list.append(wordnet_lemmatizer.lemmatize(word, pos="v"))
    lemmatized_text = " ".join(lemmatized_list)
    lemmatized_text_list.append(lemmatized_text)
    df['Content_Parsed_5'] = lemmatized_text_list
    df['Content_Parsed_6'] = df['Content_Parsed_5']
    for stop_word in stop_words:
        regex_stopword = r"\b" + stop_word + r"\b"
        df['Content_Parsed_6'] = df['Content_Parsed_6'].str.replace(regex_stc
    df = df['Content_Parsed_6']
    df = df.rename({'Content_Parsed_6': 'Content_Parsed'})

    # TF-IDF
    features = tfidf.transform(df).toarray()

    return features
```

In [427]:

```
def get_category_name(category_id):
    for category, id_ in category_codes.items():
        if id_ == category_id:
            return category
```

In [428]:
```
def predict_from_text(text):

    # Predict using the input model
    prediction_svc = svc_model.predict(create_features_from_text(text))[0]
    prediction_svc_proba = svc_model.predict_proba(create_features_from_text(

    # Return result
    category_svc = get_category_name(prediction_svc)

    print("The predicted category using the SVM model is %s." %(category_svc)
    print("The conditional probability is: %a" %(prediction_svc_proba.max()*1
```

In [429]:
```
text = """

The center-right party Ciudadanos closed a deal on Wednesday with the support

Talks in Andalusia have been ongoing since regional polls were held on Decemb

The move would see the Socialist Party lose power in the region for the first

On Thursday, Marta Bosquet of Ciudadanos was voted in as the new speaker of t

The speaker’s role in the parliament is key for the calling of an investiture

Officially, the talks as to the make up of a future government have yet to st

The speaker’s role in the parliament is key for the calling of an investiture

The PP, which was ousted from power by the PSOE in the national government in

Wednesday was a day of intense talks among the parties in a bid to find a sol

The PSOE, meanwhile, argues that having won the elections with a seven-seat l



"""
```

In [430]:
```
predict_from_text(text)
```

```
The predicted category using the SVM model is politics.
The conditional probability is: 93.21339369980114
```

In [431]:

```
# Politics

text = """Disputes have already broken out within the new political alliance

Just hours after the far-right Vox agreed to support the Popular Party (PP)'s

These early clashes suggest it could be difficult to export the model to othe

The PP and the liberal Ciudadanos have reached their own governing agreement

Ciudadanos has refused point-blank to meet with Vox representatives, but the

On Friday morning, Juan Marín of Ciudadanos said that there are no plans for

The reform party has insisted that the Vox-PP deal does not affect them at al

 Vox national leader Santiago Abascal (c) and Andalusian leader Francisco Ser
Vox national leader Santiago Abascal (c) and Andalusian leader Francisco Serr
But Vox insists on a family department, and said it will expect loyalty from

These early clashes suggest it could be difficult to export the model to othe

The PP is anxious to win back power in regions like Valencia, the Balearic Is

Parliamentary debate
The PSOE has already digested the fact that it is losing its hold on Spain's

The Socialists will not be putting forward a candidate, now that the PP nomin

The sum of the PP, Ciudadanos and Vox votes is four above the 55 required for

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is politics.
The conditional probability is: 99.43575050943763
```

In [432]:

```
# Entertainment

text = """
Cádiz is in style: it has just been included in The New York Times’ list of 5

The journalist Andrew Ferren, who wrote about Cádiz for The New York Times’ l

“Despite the fact that Cádiz was historically a major maritime link between A

Culinary delights
 Aponiente restaurant in El Puerto de Santa María.
Aponiente restaurant in El Puerto de Santa María.
Suggestions include the new Western-style gastrobar Saja River, recently open

To these suggestions, EL VIAJERO adds several of its own, including Restauran

Jerez de la Frontera and its wineries
 Bodegas Lustau, en Jerez de la Frontera (Cádiz).ampliar foto
Bodegas Lustau, en Jerez de la Frontera (Cádiz). NEIL FARRIN GETTY IMAGES
Around 36 km to the north of Cádiz lies Jerez de la Frontera, known for the f

The NMAC Montenmedio Foundation
 Vejer de la Frontera.ampliar foto
Vejer de la Frontera. GETTY IMAGES
The NMAC Montenmedio Foundation of contemporary art sits between Barbate and

EL VIAJERO expands on Ferren’s recommendations with a few of its own:

1.The Cádiz Carnival
 The Cádiz carnival.ampliar foto
The Cádiz carnival.
An unique and fun festival that takes place from February 28 to March 10. In

2. Barrio del Pópulo
 The Pópulo neighborhood.ampliar foto
The Pópulo neighborhood. RAQUEL M. CARBONELL GETTY
This is the oldest neighborhood in Cádiz and features an old Roman theater, t

3. Cádiz à la Havana
 Cathedral square in Cádiz.ampliar foto
Cathedral square in Cádiz. RAQUEL M. CARBONELL GETTY
Stroll from the colonial-style Mina Square, with its ficus and palm trees, to

4. A wealth of history
 Baelo Claudia Roman site in Tarifa (Cádiz).ampliar foto
Baelo Claudia Roman site in Tarifa (Cádiz). KEN WELSH GETTY
Standing on the frontier between two continents, the province of Cádiz has a

5. Sanlúcar de Barrameda
 Summer beach horse races in Sanlúcar de Barrameda.ampliar foto
Summer beach horse races in Sanlúcar de Barrameda. JUAN CARLOS TORO
Famous for its summer horse racing on the beach as well as for its wineries,

6. Coast and mountains
 Olvera, a white village in Cádiz.ampliar foto
Olvera, a white village in Cádiz. RUDI SEBASTIAN GETTY
```

```
Cádiz has miles of windswept beaches that make it a perfect haunt for surfers

7. The flamenco route
Located in San Fernando, the Peña Flamenca Camarón de la Isla, named after th

8. Conil de la Frontera
 The beach in Conil de la Frontera.ampliar foto
The beach in Conil de la Frontera. GETTY IMAGES
There are three national parks that stretch along Cádiz's Atlantic coast – La

9. Surfing in Tarifa
In the inlets of Los Lances and Valdevaqueros in Tarifa, wind and kitesurfers

10. The white villages
Nineteen districts in the Cádiz mountains take you through a string of white

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is entertainment.
The conditional probability is: 99.31167341445837
```

In [433]:

```
# Business

text = """
Vodafone España has informed representatives of its employees that it is putt

"In the current market climate, demand for services continues to grow exponen

Vodafone added that the current expectations of clients, "who demand an agile

As such, the company continued, it is looking to "reverse the negative trend

The operator says that it is sure it can reach a deal with labor unions so th

Vodafone has suffered a great deal in the trade war that was sparked by its r

In the first three quarters of 2018, Vodafone has lost 361,000 cellphone line

The operator executed a similar collective dismissal plan (known in Spanish a

Before the acquisition of ONO, Vodafone also executed an ERE in 2013. On that

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is business.
The conditional probability is: 93.0600852065347
```

In [434]:

```
# Tech

text = """
Elon Musk told the world in late 2017 that Tesla was taking its automotive kn

The German automaker also committed to manufacturing the truck this summer, w


While there are a few Tesla Semi prototypes on the road now, and a dozen or s

DAIMLER FIRST SHOWED OFF A PROTOTYPE IN 2015
This has left the door wide open for companies like Daimler, the parent compa

The new Cascadia is not much more advanced than the prototype was in 2015. In


The Freightliner Inspiration Truck at the event in 2015.
But the new Cascadia is doing this with a limited set of sensors. There's a f

This helps keep costs down, but means the technology is more in line with wha

DAIMLER'S TRUCK HAS MORE IN COMMON WITH NISSAN'S PROPILOT SYSTEM THAN TESLA'S
Keeping with a theme of less is more, there's also no camera-based monitoring

A sensor in the steering column measures resistance applied to the steering w



The new Cascadia is a far cry from a fully autonomous truck, but based on my

A Daimler representative also told me that, while lane centering is on, the d

RELATED

This is what it's like to ride in Daimler's self-driving semi truck
Daimler promised some other modern technologies are coming the new Cascadia,

The Cascadia won't be as stuffed with tech as the Tesla Semi, nor is it as sl
"""

predict_from_text(text)
```

```
The predicted category using the SVM model is tech.
The conditional probability is: 98.17546586390013
```

In [435]:

```
# Sports

text = """
Spain has agreed to host the soccer final of the Copa Libertadores between Ar

The final in Madrid is a punch in the soul to all fans of soccer in Argentina

ONLINE SPORTS DAILY OLE

The final was set to take place in Argentina but was suspended twice after fa

In view of the insecurity, the South American Football Confederation (Conmebo

 Embedded video

Sebastián Lisiecki
@sebalisiecki
 Así fue la llegada de Boca al Monumental. Pésimo la seguridad q los mete ent

575
7:23 PM - Nov 24, 2018
637 people are talking about this
Twitter Ads info and privacy
This was how Boca arrived at Monumental stadium. The security that got betwee

This is the first time a Copa Libertadores game has been played outside the A

But the feeling in Argentina has been less optimistic. The national newspaper

Security risk
In a message on Twitter, Sánchez promised that "security forces have extensiv

River and Boca have a long-standing rivalry fueled largely by the class divid

Scheduling issues
The final will take place on Sunday, December 9, on the final day of a three-

 Conmebol president Alejandro Domínguez on Tuesday.
Conmebol president Alejandro Domínguez on Tuesday.
Many details about the game have yet to be revealed, including how tickets wi

Conmebol and soccer club representatives began considering destinations for t



"""

predict_from_text(text)
```

```
The predicted category using the SVM model is sport.
The conditional probability is: 75.68806067700831
```

In [436]:

```
# Weather

text = """
A polar air mass that entered the Iberian peninsula on Wednesday has already

“An episode of intense cold” is forecast for Friday, when the mercury will co

Elsewhere, weather stations have recorded -8.2ºC in La Molina (Girona), at an

 Almería has rolled out vehicles to deal with wintry road conditions.
Almería has rolled out vehicles to deal with wintry road conditions. DIPUTACI
Aemet spokesman Rubén del Campo said that the cold spell is not out of the or

Temperatures have already dipped between six and eight degrees in a matter of

Temperatures on Friday and Saturday will be “very cold, with lows of five to

No snow
However, little to no snow is expected “not for lack of cold, but for lack of

Alerts are in place in Almería, Granada, Jaén, Aragón, Cantabria, Castilla-La

On Saturday, the orange warnings will extend to Córdoba, Salamanca, Valladoli


"""

predict_from_text(text)
```

```
The predicted category using the SVM model is business.
The conditional probability is: 62.95086242483375
```

In [437]:

```
# Health

text = """
The obesity epidemic has been on the rise for years, with cases nearly tripli

An investigation by the Mar de Barcelona hospital has found that 80% of men a

Being overweight can mean a higher risk of suffering a number of diseases, in

The study, published in the Spanish Cardiology Magazine, points out that this

The issue, the experts state, is not an esthetic one, but rather a question o

Researchers at the Barcelona hospital revised all of the scientific literatur

There are currently 25 million people with excess weight, three million more

DR ALBERT GODAY, AUTHOR OF THE STUDY

“There are currently 25 million people with excess weight, three million more

“In men, excess weight is more usual up to the age of 50,” explains Goday. “F

The experts argue that any weight loss, no matter how small, reduces the risk

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is tech.
The conditional probability is: 40.79994044584591
```

In [438]:

```
# Animal abuse

text = """
Spain’s animal rights party PACMA posted a 38-second video on Twitter on Frid

“Hunters shut what appears to be a fox in a cage and let it out only to peppe

 Video insertado

PACMA
✓
@PartidoPACMA
 Cazadores enjaulan a lo que parece ser un zorro y lo liberan solo para acrib

En realidad, son peligrosos psicópatas con escopeta y permiso de amas. #YoNoD

4.188
10:43 - 4 ene. 2019
7.443 personas están hablando de esto
Información y privacidad de Twitter Ads
At the start of the video, a man teases the caged animal with a stick. When t

The release of the video, which has had 255,000 views, coincided with the lau

As it notes on its website, PACMA is the only political group that opposes hu

No animal should die under fire. We will fight tirelessly until hunting becom

PACMA

The animal rights group is preparing a report to send to the regional governm

Last month, a Spanish hunter who was filmed while he chased and tortured a fo

And in November, animal rights groups and political parties reacted with indi

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is entertainment.
The conditional probability is: 50.955623421406294
```

Part 2

After successfully implementing their code. Try to gather data from an online URL related to autonomous cars (your choice but a long article) Use all techniques covered in the above code on the dataset that you have just created.

#00,01-Data Creation,02-Exploratory Data Analysis

In [439]:
```
df_path = ""
#https://storage.googleapis.com/dataset-uploader/bbc/bbc-text.csv
df_path2 = df_path + 'Part2_dataset.csv'
df = pd.read_csv(df_path2)
df.head()
```

Out[439]:

	category	text
0	tech	tv future in the hands of viewers with home th...
1	business	worldcom boss left books alone former worldc...
2	sport	tigers wary of farrell gamble leicester say ...
3	sport	yeading face newcastle in fa cup premiership s...
4	entertainment	ocean s twelve raids box office ocean s twelve...

##Number of articles in each category

In [440]:
```
bars = alt.Chart(df).mark_bar(size=50).encode(
    x=alt.X("category"),
    y=alt.Y("count():Q", axis=alt.Axis(title='Number of articles')),
    tooltip=[alt.Tooltip('count()', title='Number of articles'), 'category'],
    color='category'

)

text = bars.mark_text(
    align='center',
    baseline='bottom',
).encode(
    text='count()'
)

(bars + text).interactive().properties(
    height=300,
    width=700,
    title = "Number of articles in each category",
)
```

```
<vega.vegalite.VegaLite at 0x7f00a9ee0d50>
```

Out[440]:

##% of articles in each category

In [441]:

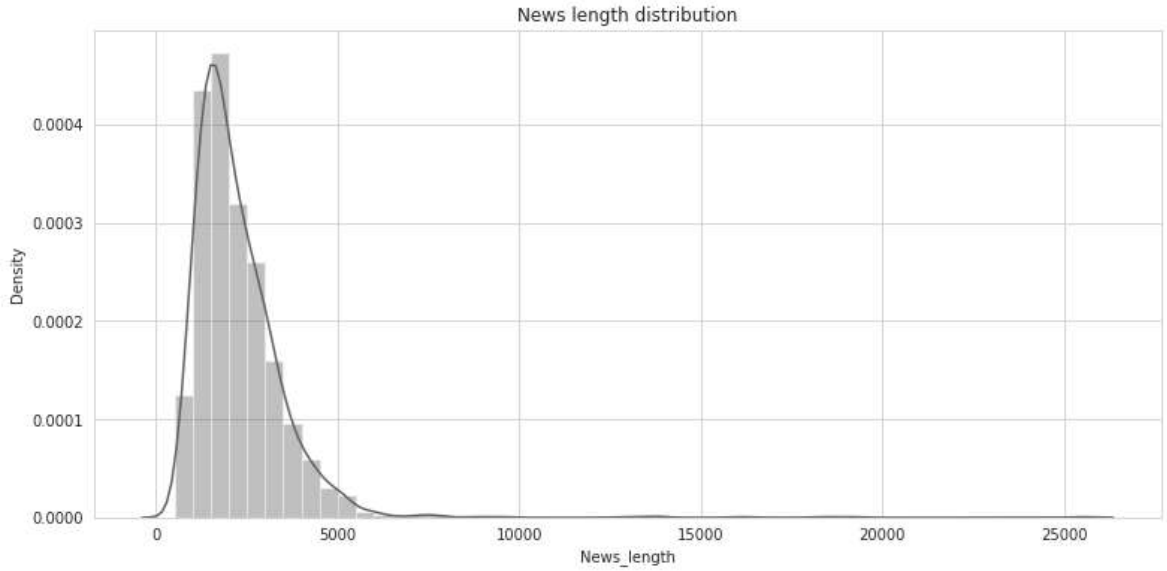
```
df['id'] = 1
df2 = pd.DataFrame(df.groupby('category').count()['id']).reset_index()

bars = alt.Chart(df2).mark_bar(size=50).encode(
    x=alt.X('category'),
    y=alt.Y('PercentOfTotal:Q', axis=alt.Axis(format='.0%', title='% of Artic
    color='category'
).transform_window(
    TotalArticles='sum(id)',
    frame=[None, None]
).transform_calculate(
    PercentOfTotal="datum.id / datum.TotalArticles"
)

text = bars.mark_text(
    align='center',
    baseline='bottom',
    #dx=5  # Nudges text to right so it doesn't appear on top of the bar
).encode(
    text=alt.Text('PercentOfTotal:Q', format='.1%')
)

(bars + text).interactive().properties(
    height=300,
    width=700,
    title = "% of articles in each category",
)
```

```
<vega.vegalite.VegaLite at 0x7f00ad1e4a90>
```

Out[441]:

##News length by category

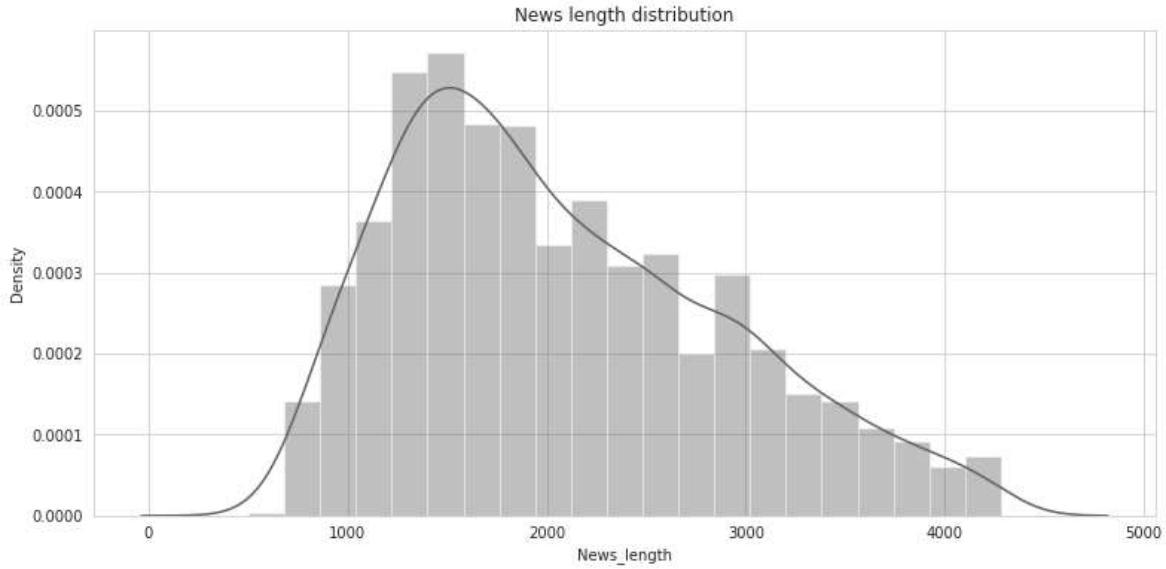
In [442]:
```
df['News_length'] = df['text'].str.len()

plt.figure(figsize=(9.5,7))
sns.distplot(df['News_length']).set_title('News length distribution');
```



In [443]:
```
df['News_length'].describe()
```

Out[443]:
```
count     2225.00000
mean      2262.93618
std       1364.10253
min        501.00000
25%       1446.00000
50%       1965.00000
75%       2802.00000
max      25483.00000
Name: News_length, dtype: float64
```

In [444]:
```
quantile_95 = df['News_length'].quantile(0.95)
df_95 = df[df['News_length'] < quantile_95]
plt.figure(figsize=(9.5,7))
sns.distplot(df_95['News_length']).set_title('News length distribution');
```



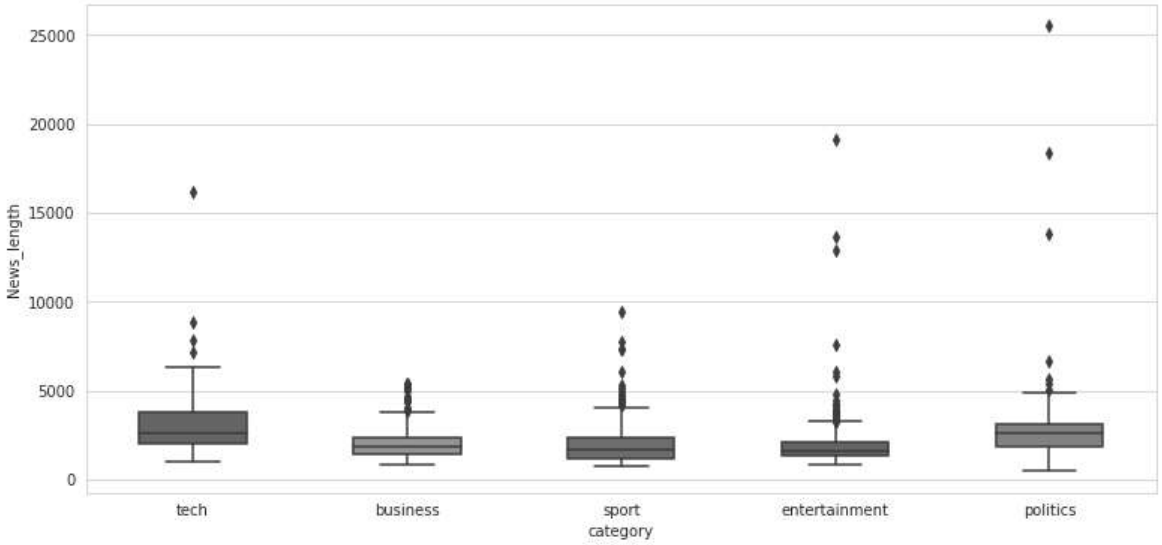
In [445]:
```
df_more10k = df[df['News_length'] > 10000]
len(df_more10k)
```

Out[445]: 7

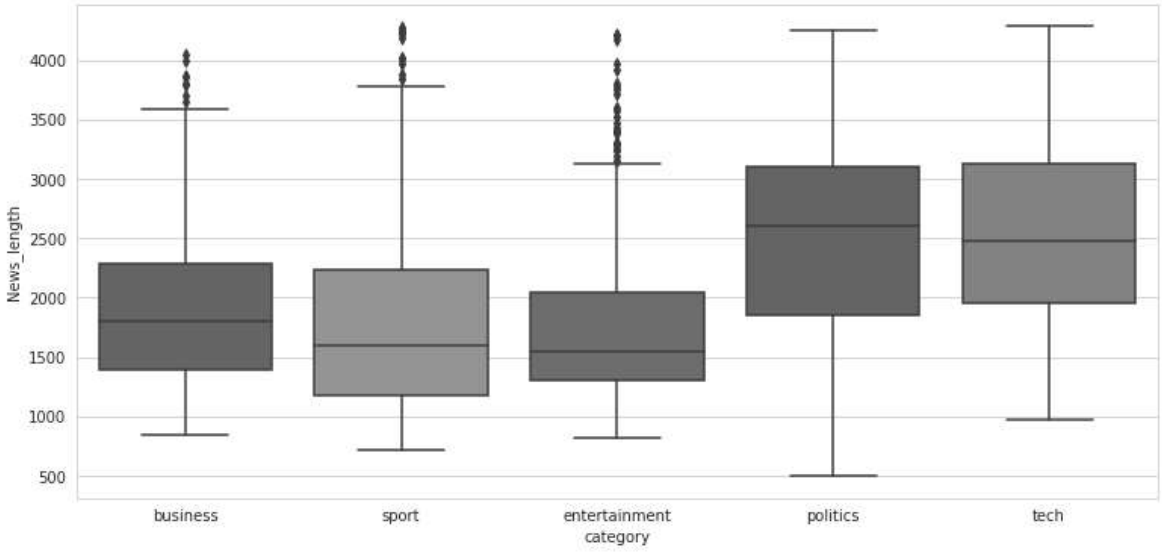
In [446]:
```
df_more10k['text'].iloc[0]
```

```
n 25 august  1938. the mortal threat back then was a scruffy little aust
rian called adolf hitler. a week after my first birthday  the threat had
become reality. we were at war. my father wore a uniform for five years.
after 1945 we yearned for peace at last. but in 1946 winston churchill t
old us - from the baltic to the adriatic an iron curtain has descended a
cross europe. behind the iron curtain  another genocidal psychopath  ano
ther threat. josef stalin triggered the cold war  with the berlin blocka
de in 1948. my whole generation was blighted by it.  we were threatened
by the nuclear holocaust  the nuclear wind  the nuclear winter. we built
shelters that would have sheltered nothing. we spent our treasure on wea
pons instead of hospitals. we took silly precautions. some fought it; so
me marched futilely against it. some pretended it was not there. the col
d war lasted 43 years  but we remained a parliamentary democracy. by the
early seventies it was terrorism as well. al fatah  black september  red
brigades  but most of all for us the ira and the inla. thirty more year
s; 300 policemen and women  over 600 soldiers  more than 3 000 civilians
dead  but we won because even ira bombs could not force us to become a t
yranny. that was why the tyrants lost. civil rights were infringed as li
ttle as humanly possible. evidence had to be taken in secret to protect
covert sources: yes  and one judge  no-jury courts had to be instituted
```

In [447]:
```
plt.figure(figsize=(9.5,7))
sns.boxplot(data=df, x='category', y='News_length', width=.5);
```



In [448]:
```
plt.figure(figsize=(9.5,7))
sns.boxplot(data=df_95, x='category', y='News_length');
```



In [449]:
```
with open('Part2_dataset.pickle', 'wb') as output:
    pickle.dump(df, output)
```

03

In [450]:
```
path_df = "Part2_dataset.pickle"

with open(path_df, 'rb') as data:
    df = pickle.load(data)

df.head()
```

Out[450]:

	category	text	id	News_length
0	tech	tv future in the hands of viewers with home th...	1	4333
1	business	worldcom boss left books alone former worldc...	1	1842
2	sport	tigers wary of farrell gamble leicester say ...	1	1342
3	sport	yeading face newcastle in fa cup premiership s...	1	2176
4	entertainment	ocean s twelve raids box office ocean s twelve...	1	1579

In [451]:
```
df.loc[1]['text']
```

Out[451]:
```
'worldcom boss  left books alone  former worldcom boss bernie ebbers  who i
s accused of overseeing an $11bn (£5.8bn) fraud  never made accounting deci
sions  a witness has told jurors.  david myers made the comments under ques
tioning by defence lawyers who have been arguing that mr ebbers was not res
ponsible for worldcom s problems. the phone company collapsed in 2002 and p
rosecutors claim that losses were hidden to protect the firm s shares. mr m
yers has already pleaded guilty to fraud and is assisting prosecutors.  on
monday  defence lawyer reid weingarten tried to distance his client from th
e allegations. during cross examination  he asked mr myers if he ever knew
mr ebbers  make an accounting decision  .  not that i am aware of  mr myer
s replied.  did you ever know mr ebbers to make an accounting entry into wo
rldcom books  mr weingarten pressed.  no  replied the witness. mr myers h
as admitted that he ordered false accounting entries at the request of form
er worldcom chief financial officer scott sullivan. defence lawyers have be
en trying to paint mr sullivan  who has admitted fraud and will testify lat
er in the trial  as the mastermind behind worldcom s accounting house of ca
rds.  mr ebbers  team  meanwhile  are looking to portray him as an affable
boss  who by his own admission is more pe graduate than economist. whatever
his abilities  mr ebbers transformed worldcom from a relative unknown into
a $160bn telecoms giant and investor darling of the late 1990s. worldcom s
problems mounted  however  as competition increased and the telecoms boom p
etered out. when the firm finally collapsed  shareholders lost about $180bn
and 20 000 workers lost their jobs. mr ebbers  trial is expected to last tw
o months and if found guilty the former ceo faces a substantial jail senten
ce. he has firmly declared his innocence.'
```

##Text Cleaning

###Special Character Cleaning

In [452]:
```
# \r and \n
df['Content_Parsed_1'] = df['text'].str.replace("\r", " ")
df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace("\n", " ")
df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace("    ", " ")
# " when quoting text
df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace('"', '')
text = "Mr Greenspan\'s"
text
```

Out[452]: "Mr Greenspan's"

Upcase/downcase

In [453]:
```
# Lowercasing the text
df['Content_Parsed_2'] = df['Content_Parsed_1'].str.lower()
```

Punctuation signs

In [454]:
```
punctuation_signs = list("?:!.,;")
df['Content_Parsed_3'] = df['Content_Parsed_2']

for punct_sign in punctuation_signs:
    df['Content_Parsed_3'] = df['Content_Parsed_3'].str.replace(punct_sign, '
```

Possessive Pronouns

In [455]:
```
df['Content_Parsed_4'] = df['Content_Parsed_3'].str.replace("'s", "")
```

Stemming and Lemmatization

In [456]:
```
# Saving the lemmatizer into an object
wordnet_lemmatizer = WordNetLemmatizer()

nrows = len(df)
lemmatized_text_list = []

for row in range(0, nrows):

    # Create an empty list containing lemmatized words
    lemmatized_list = []

    # Save the text and its words into an object
    text = df.loc[row]['Content_Parsed_4']
    text_words = text.split(" ")

    # Iterate through every word to lemmatize
    for word in text_words:
        lemmatized_list.append(wordnet_lemmatizer.lemmatize(word, pos="v"))

    # Join the list
    lemmatized_text = " ".join(lemmatized_list)

    # Append to the list containing the texts
    lemmatized_text_list.append(lemmatized_text)

df['Content_Parsed_5'] = lemmatized_text_list
```

Stop words

In [457]:
```
# Loading the stop words in english
stop_words = list(stopwords.words('english'))
stop_words[0:10]
```

Out[457]:
```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]
```

In [458]:
```
example = "me eating a meal"
word = "me"

# The regular expression is:
regex = r"\b" + word + r"\b"  # we need to build it like that to work properl

re.sub(regex, "StopWord", example)
```

Out[458]:
```
'StopWord eating a meal'
```

In [459]:

```
df['Content_Parsed_6'] = df['Content_Parsed_5']

for stop_word in stop_words:

    regex_stopword = r"\b" + stop_word + r"\b"
    df['Content_Parsed_6'] = df['Content_Parsed_6'].str.replace(regex_stopwor
```

Results of parsing

In [460]:
```
df.loc[5]['text']
```

Out[460]:
```
'howard hits back at mongrel jibe michael howard has said a claim by peter
hain that the tory leader is acting like an  attack mongrel  shows labour i
s  rattled  by the opposition.  in an upbeat speech to his party s spring c
onference in brighton  he said labour s campaigning tactics proved the tori
es were hitting home. mr hain made the claim about tory tactics in the anti
-terror bill debate.  something tells me that someone  somewhere out there
is just a little bit rattled  mr howard said. mr hain  leader of the commo
ns  told bbc radio four s today programme that mr howard s stance on the go
vernment s anti-terrorism legislation was putting the country at risk. he t
hen accused the tory leader of behaving like an  attack mongrel  and  playi
ng opposition for opposition sake .  mr howard told his party that labour w
ould  do anything  say anything  claim anything to cling on to office at al
l costs .  so far this year they have compared me to fagin  to shylock and
to a flying pig. this morning peter hain even called me a mongrel.  i don t
know about you  but something tells me that someone  somewhere out there is
just a little bit rattled.  environment secretary margaret beckett rejected
mr howard s comment  telling radio 4 s pm programme that labour was not  ra
ttled .  we have a very real duty to try to get people to focus on michael
howard s record  what the proposals are that he is trying to put forward to
the country and also the many examples we are seeing now of what we believe
is really poor judgement on his behalf.  mr howard said tory policies on s
chools  taxes  immigration and crime were striking a chord with voters.  s
ince the beginning of this year - election year - we ve been making the pol
itical weather  he told the party conference. mr howard denied he had been
playing politics  by raising the case of margaret dixon  whose operation ha
d been cancelled seven times  which grabbed headlines for the party two wee
ks ago. and he hit back at labour claims he had used mrs dixon as a  human
shield .  she s not a human shield mr blair  she s a human being.  mr howar
d said his party plans for immigration quotas  which have also been the foc
us of much media coverage  were not  racist  - just  common sense .  he ple
dged cleaner hospitals and better school discipline  with a promise to get
rid of  political correctness  in the national curriculum and give everyone
to the same chance of a  decent  state education as he had.  i come from an
ordinary family. if the teenage michael howard were applying to cambridge t
oday  gordon brown would love me.  and he stressed his party s commitment
to cut taxes and red tape and increase the basic state pension in line with
earnings. he finished with a personal appeal to party activists to go out a
nd win the next election.  one day you will be able to tell your children a
nd grandchildren as i will tell mine  i was there. i did my bit. i played
my part. i helped to win that famous election - the election that transform
ed our country for the better .  labour election co-ordinator alan milburn
said:  michael howard s speech today confirms what we have always said - th
at his only strategy is opportunism but he has no forward vision for the co
untry. in reference to the appearance of mr howard s family on the conferen
ce stage with him  mr milburn said:  michael howard is perfectly entitled t
o pose with his family today.  but it is the hard working families across b
ritain that will be damaged by his plan to cut £35bn from public spending.'
```

In [461]:
```
df.loc[5]['Content_Parsed_1']
```

Out[461]:
```
'howard hits back at mongrel jibe michael howard has said a claim by peter
hain that the tory leader is acting like an  attack mongrel  shows labour i
s  rattled  by the opposition.  in an upbeat speech to his party s spring c
onference in brighton  he said labour s campaigning tactics proved the tori
es were hitting home. mr hain made the claim about tory tactics in the anti
-terror bill debate.  something tells me that someone  somewhere out there
is just a little bit rattled  mr howard said. mr hain  leader of the commo
ns  told bbc radio four s today programme that mr howard s stance on the go
vernment s anti-terrorism legislation was putting the country at risk. he t
hen accused the tory leader of behaving like an  attack mongrel  and  playi
ng opposition for opposition sake .  mr howard told his party that labour w
ould  do anything  say anything  claim anything to cling on to office at al
l costs .  so far this year they have compared me to fagin  to shylock and
to a flying pig. this morning peter hain even called me a mongrel.  i don t
know about you  but something tells me that someone  somewhere out there is
just a little bit rattled.  environment secretary margaret beckett rejected
mr howard s comment  telling radio 4 s pm programme that labour was not  ra
ttled .  we have a very real duty to try to get people to focus on michael
howard s record  what the proposals are that he is trying to put forward to
the country and also the many examples we are seeing now of what we believe
is really poor judgement on his behalf.  mr howard said tory policies on s
chools  taxes  immigration and crime were striking a chord with voters.  s
ince the beginning of this year - election year - we ve been making the pol
itical weather  he told the party conference. mr howard denied he had been
playing politics  by raising the case of margaret dixon  whose operation ha
d been cancelled seven times  which grabbed headlines for the party two wee
ks ago. and he hit back at labour claims he had used mrs dixon as a  human
shield .  she s not a human shield mr blair  she s a human being.  mr howar
d said his party plans for immigration quotas  which have also been the foc
us of much media coverage  were not  racist  - just  common sense .  he ple
dged cleaner hospitals and better school discipline  with a promise to get
rid of  political correctness  in the national curriculum and give everyone
to the same chance of a  decent  state education as he had.  i come from an
ordinary family. if the teenage michael howard were applying to cambridge t
oday  gordon brown would love me.  and he stressed his party s commitment
to cut taxes and red tape and increase the basic state pension in line with
earnings. he finished with a personal appeal to party activists to go out a
nd win the next election.  one day you will be able to tell your children a
nd grandchildren as i will tell mine  i was there. i did my bit. i played
my part. i helped to win that famous election - the election that transform
ed our country for the better .  labour election co-ordinator alan milburn
said:  michael howard s speech today confirms what we have always said - th
at his only strategy is opportunism but he has no forward vision for the co
untry. in reference to the appearance of mr howard s family on the conferen
ce stage with him  mr milburn said:  michael howard is perfectly entitled t
o pose with his family today.  but it is the hard working families across b
ritain that will be damaged by his plan to cut £35bn from public spending.'
```

In [462]:
```
df.loc[5]['Content_Parsed_2']
```

Out[462]:
```
'howard hits back at mongrel jibe michael howard has said a claim by peter
hain that the tory leader is acting like an  attack mongrel  shows labour i
s  rattled  by the opposition.  in an upbeat speech to his party s spring c
onference in brighton  he said labour s campaigning tactics proved the tori
es were hitting home. mr hain made the claim about tory tactics in the anti
-terror bill debate.  something tells me that someone  somewhere out there
is just a little bit rattled  mr howard said. mr hain  leader of the commo
ns  told bbc radio four s today programme that mr howard s stance on the go
vernment s anti-terrorism legislation was putting the country at risk. he t
hen accused the tory leader of behaving like an  attack mongrel  and  playi
ng opposition for opposition sake .  mr howard told his party that labour w
ould  do anything  say anything  claim anything to cling on to office at al
l costs .  so far this year they have compared me to fagin  to shylock and
to a flying pig. this morning peter hain even called me a mongrel.  i don t
know about you  but something tells me that someone  somewhere out there is
just a little bit rattled.  environment secretary margaret beckett rejected
mr howard s comment  telling radio 4 s pm programme that labour was not  ra
ttled .  we have a very real duty to try to get people to focus on michael
howard s record  what the proposals are that he is trying to put forward to
the country and also the many examples we are seeing now of what we believe
is really poor judgement on his behalf.  mr howard said tory policies on s
chools  taxes  immigration and crime were striking a chord with voters.  s
ince the beginning of this year - election year - we ve been making the pol
itical weather  he told the party conference. mr howard denied he had been
playing politics  by raising the case of margaret dixon  whose operation ha
d been cancelled seven times  which grabbed headlines for the party two wee
ks ago. and he hit back at labour claims he had used mrs dixon as a  human
shield .  she s not a human shield mr blair  she s a human being.  mr howar
d said his party plans for immigration quotas  which have also been the foc
us of much media coverage  were not  racist  - just  common sense .  he ple
dged cleaner hospitals and better school discipline  with a promise to get
rid of  political correctness  in the national curriculum and give everyone
to the same chance of a  decent  state education as he had.  i come from an
ordinary family. if the teenage michael howard were applying to cambridge t
oday  gordon brown would love me.  and he stressed his party s commitment
to cut taxes and red tape and increase the basic state pension in line with
earnings. he finished with a personal appeal to party activists to go out a
nd win the next election.  one day you will be able to tell your children a
nd grandchildren as i will tell mine  i was there. i did my bit. i played
my part. i helped to win that famous election - the election that transform
ed our country for the better .  labour election co-ordinator alan milburn
said:  michael howard s speech today confirms what we have always said - th
at his only strategy is opportunism but he has no forward vision for the co
untry. in reference to the appearance of mr howard s family on the conferen
ce stage with him  mr milburn said:  michael howard is perfectly entitled t
o pose with his family today.  but it is the hard working families across b
ritain that will be damaged by his plan to cut £35bn from public spending.'
```

In [463]:
```
df.loc[5]['Content_Parsed_3']
```

Out[463]:
```
'howard hits back at mongrel jibe michael howard has said a claim by peter
hain that the tory leader is acting like an  attack mongrel  shows labour i
s  rattled  by the opposition  in an upbeat speech to his party s spring co
nference in brighton  he said labour s campaigning tactics proved the torie
s were hitting home mr hain made the claim about tory tactics in the anti-t
error bill debate  something tells me that someone  somewhere out there is
just a little bit rattled   mr howard said mr hain  leader of the commons
told bbc radio four s today programme that mr howard s stance on the govern
ment s anti-terrorism legislation was putting the country at risk he then a
ccused the tory leader of behaving like an  attack mongrel  and  playing op
position for opposition sake   mr howard told his party that labour would
do anything  say anything  claim anything to cling on to office at all cost
s   so far this year they have compared me to fagin  to shylock and to a fl
ying pig this morning peter hain even called me a mongrel  i don t know abo
ut you  but something tells me that someone  somewhere out there is just a
little bit rattled  environment secretary margaret beckett rejected mr howa
rd s comment  telling radio 4 s pm programme that labour was not  rattled
we have a very real duty to try to get people to focus on michael howard s
record  what the proposals are that he is trying to put forward to the coun
try and also the many examples we are seeing now of what we believe is real
ly poor judgement on his behalf   mr howard said tory policies on schools
taxes  immigration and crime were striking a chord with voters   since the
beginning of this year - election year - we ve been making the political we
ather   he told the party conference mr howard denied he had been  playing
politics  by raising the case of margaret dixon  whose operation had been c
ancelled seven times  which grabbed headlines for the party two weeks ago a
nd he hit back at labour claims he had used mrs dixon as a  human shield
she s not a human shield mr blair  she s a human being  mr howard said his
party plans for immigration quotas  which have also been the focus of much
media coverage  were not  racist  - just  common sense   he pledged cleaner
hospitals and better school discipline  with a promise to get rid of  polit
ical correctness  in the national curriculum and give everyone to the same
chance of a  decent  state education as he had  i come from an ordinary fam
ily if the teenage michael howard were applying to cambridge today  gordon
brown would love me   and he stressed his party s commitment to cut taxes a
nd red tape and increase the basic state pension in line with earnings he f
inished with a personal appeal to party activists to go out and win the nex
t election  one day you will be able to tell your children and grandchildre
n as i will tell mine   i was there i did my bit i played my part i helped
to win that famous election - the election that transformed our country for
the better   labour election co-ordinator alan milburn said  michael howard
s speech today confirms what we have always said - that his only strategy i
s opportunism but he has no forward vision for the country in reference to
the appearance of mr howard s family on the conference stage with him  mr m
ilburn said  michael howard is perfectly entitled to pose with his family t
oday  but it is the hard working families across britain that will be damag
ed by his plan to cut £35bn from public spending'
```

In [464]:
```
df.loc[5]['Content_Parsed_4']
```

Out[464]:
```
'howard hits back at mongrel jibe michael howard has said a claim by peter
hain that the tory leader is acting like an  attack mongrel  shows labour i
s  rattled  by the opposition  in an upbeat speech to his party s spring co
nference in brighton  he said labour s campaigning tactics proved the torie
s were hitting home mr hain made the claim about tory tactics in the anti-t
error bill debate  something tells me that someone  somewhere out there is
just a little bit rattled   mr howard said mr hain  leader of the commons
told bbc radio four s today programme that mr howard s stance on the govern
ment s anti-terrorism legislation was putting the country at risk he then a
ccused the tory leader of behaving like an  attack mongrel  and  playing op
position for opposition sake   mr howard told his party that labour would
do anything  say anything  claim anything to cling on to office at all cost
s   so far this year they have compared me to fagin  to shylock and to a fl
ying pig this morning peter hain even called me a mongrel  i don t know abo
ut you  but something tells me that someone  somewhere out there is just a
little bit rattled  environment secretary margaret beckett rejected mr howa
rd s comment  telling radio 4 s pm programme that labour was not  rattled
we have a very real duty to try to get people to focus on michael howard s
record  what the proposals are that he is trying to put forward to the coun
try and also the many examples we are seeing now of what we believe is real
ly poor judgement on his behalf   mr howard said tory policies on schools
taxes  immigration and crime were striking a chord with voters   since the
beginning of this year - election year - we ve been making the political we
ather   he told the party conference mr howard denied he had been  playing
politics  by raising the case of margaret dixon  whose operation had been c
ancelled seven times  which grabbed headlines for the party two weeks ago a
nd he hit back at labour claims he had used mrs dixon as a  human shield
she s not a human shield mr blair  she s a human being  mr howard said his
party plans for immigration quotas  which have also been the focus of much
media coverage  were not  racist  - just  common sense   he pledged cleaner
hospitals and better school discipline  with a promise to get rid of  polit
ical correctness  in the national curriculum and give everyone to the same
chance of a  decent  state education as he had  i come from an ordinary fam
ily if the teenage michael howard were applying to cambridge today  gordon
brown would love me   and he stressed his party s commitment to cut taxes a
nd red tape and increase the basic state pension in line with earnings he f
inished with a personal appeal to party activists to go out and win the nex
t election  one day you will be able to tell your children and grandchildre
n as i will tell mine   i was there i did my bit i played my part i helped
to win that famous election - the election that transformed our country for
the better   labour election co-ordinator alan milburn said  michael howard
s speech today confirms what we have always said - that his only strategy i
s opportunism but he has no forward vision for the country in reference to
the appearance of mr howard s family on the conference stage with him  mr m
ilburn said  michael howard is perfectly entitled to pose with his family t
oday  but it is the hard working families across britain that will be damag
ed by his plan to cut £35bn from public spending'
```

In [465]:
```
df.loc[5]['Content_Parsed_5']
```

Out[465]:
```
'howard hit back at mongrel jibe michael howard have say a claim by peter h
ain that the tory leader be act like an  attack mongrel  show labour be  ra
ttle  by the opposition  in an upbeat speech to his party s spring conferen
ce in brighton  he say labour s campaign tactics prove the tories be hit ho
me mr hain make the claim about tory tactics in the anti-terror bill debate
something tell me that someone  somewhere out there be just a little bite r
attle   mr howard say mr hain  leader of the commons  tell bbc radio four s
today programme that mr howard s stance on the government s anti-terrorism
legislation be put the country at risk he then accuse the tory leader of be
have like an  attack mongrel  and  play opposition for opposition sake   mr
howard tell his party that labour would  do anything  say anything  claim a
nything to cling on to office at all cost   so far this year they have comp
are me to fagin  to shylock and to a fly pig this morning peter hain even c
all me a mongrel  i don t know about you  but something tell me that someon
e  somewhere out there be just a little bite rattle  environment secretary
margaret beckett reject mr howard s comment  tell radio 4 s pm programme th
at labour be not  rattle   we have a very real duty to try to get people to
focus on michael howard s record  what the proposals be that he be try to p
ut forward to the country and also the many examples we be see now of what
we believe be really poor judgement on his behalf   mr howard say tory poli
cies on school  tax  immigration and crime be strike a chord with voters
since the begin of this year - election year - we ve be make the political
weather   he tell the party conference mr howard deny he have be  play poli
tics  by raise the case of margaret dixon  whose operation have be cancel s
even time  which grab headline for the party two weeks ago and he hit back
at labour claim he have use mrs dixon as a  human shield   she s not a huma
n shield mr blair  she s a human be  mr howard say his party plan for immig
ration quotas  which have also be the focus of much media coverage  be not
racist  - just  common sense   he pledge cleaner hospitals and better schoo
l discipline  with a promise to get rid of  political correctness  in the n
ational curriculum and give everyone to the same chance of a  decent  state
education as he have  i come from an ordinary family if the teenage michael
howard be apply to cambridge today  gordon brown would love me   and he str
ess his party s commitment to cut tax and red tape and increase the basic s
tate pension in line with earn he finish with a personal appeal to party ac
tivists to go out and win the next election  one day you will be able to te
ll your children and grandchildren as i will tell mine   i be there i do my
bite i play my part i help to win that famous election - the election that
transform our country for the better   labour election co-ordinator alan mi
lburn say  michael howard s speech today confirm what we have always say -
that his only strategy be opportunism but he have no forward vision for the
country in reference to the appearance of mr howard s family on the confere
nce stage with him  mr milburn say  michael howard be perfectly entitle to
pose with his family today  but it be the hard work families across britain
that will be damage by his plan to cut £35bn from public spend'
```

In [466]:
```
df.loc[5]['Content_Parsed_6']
```

Out[466]:
```
'howard hit back  mongrel jibe michael howard  say  claim  peter hain  tor
y leader  act like  attack mongrel  show labour  rattle  opposition
upbeat speech  party  spring conference  brighton  say labour  campaign t
actics prove  tories  hit home mr hain make  claim  tory tactics  anti-ter
ror bill debate  something tell  someone  somewhere  little bite rattl
e  mr howard say mr hain  leader  commons  tell bbc radio four  today pro
gramme  mr howard  stance  government  anti-terrorism legislation  put  co
untry  risk  accuse  tory leader  behave like  attack mongrel  play opp
osition  opposition sake  mr howard tell  party  labour would  anything
say anything  claim anything  cling  office  cost  far  year  compare
fagin  shylock  fly pig  morning peter hain even call  mongrel  know
something tell  someone  somewhere  little bite rattle  environment se
cretary margaret beckett reject mr howard  comment  tell radio 4  pm progra
mme  labour  rattle  real duty  try  get people  focus  michael howa
rd  record  proposals  try  put forward  country  also  many examples
see  believe  really poor judgement  behalf  mr howard say tory polici
es  school  tax  immigration  crime  strike  chord  voters  since  begin
year - election year -  make  political weather  tell  party conference
mr howard deny  play politics  raise  case  margaret dixon  whose opera
tion  cancel seven time  grab headline  party two weeks ago  hit back
labour claim  use mrs dixon  human shield  human shield mr blair
human  mr howard say  party plan  immigration quotas  also  focus  much
media coverage  racist -  common sense  pledge cleaner hospitals  be
tter school discipline  promise  get rid  political correctness  natio
nal curriculum  give everyone  chance  decent  state education  com
e  ordinary family  teenage michael howard  apply  cambridge today  gordo
n brown would love  stress  party  commitment  cut tax  red tape  incre
ase  basic state pension  line  earn  finish  personal appeal  party activ
ists  go  win  next election  one day  able  tell  children  grandchildr
en  tell mine  bite  play  part  help  win  famous election -  ele
ction  transform  country  better  labour election co-ordinator alan milb
urn say  michael howard  speech today confirm  always say -  strategy
opportunism  forward vision  country  reference  appearance  mr howard
family  conference stage  mr milburn say  michael howard  perfectly enti
tle  pose  family today  hard work families across britain  damage
plan  cut £35bn  public spend'
```

In [467]:
```
df.head(1)
```

Out[467]:

	category	text	id	News_length	Content_Parsed_1	Content_Parsed_2	Content_Parsed_3
0	tech	tv future in the hands of viewers with home th...	1	4333	tv future in the hands of viewers with home th...	tv future in the hands of viewers with home th...	tv future in the hands of viewers with home th..

In [468]:
```
list_columns = ["category", "text", "Content_Parsed_6"]
df = df[list_columns]

df = df.rename(columns={'Content_Parsed_6': 'Content_Parsed'})

df.head()
```

Out[468]:

	category	text	Content_Parsed
0	tech	tv future in the hands of viewers with home th...	tv future hand viewers home theatre system...
1	business	worldcom boss left books alone former worldc...	worldcom boss leave book alone former worldc...
2	sport	tigers wary of farrell gamble leicester say ...	tigers wary farrell gamble leicester say ...
3	sport	yeading face newcastle in fa cup premiership s...	yeading face newcastle fa cup premiership sid...
4	entertainment	ocean s twelve raids box office ocean s twelve...	ocean twelve raid box office ocean twelve ...

##Label Coding

In [469]:
```
category_codes = {
    'business': 0,
    'entertainment': 1,
    'politics': 2,
    'sport': 3,
    'tech': 4
}

# Category mapping
df['Category_Code'] = df['category']
df = df.replace({'Category_Code':category_codes})

df.head()
```

Out[469]:

	category	text	Content_Parsed	Category_Code
0	tech	tv future in the hands of viewers with home th...	tv future hand viewers home theatre system...	4
1	business	worldcom boss left books alone former worldc...	worldcom boss leave book alone former worldc...	0
2	sport	tigers wary of farrell gamble leicester say ...	tigers wary farrell gamble leicester say ...	3
3	sport	yeading face newcastle in fa cup premiership s...	yeading face newcastle fa cup premiership sid...	3
4	entertainment	ocean s twelve raids box office ocean s twelve...	ocean twelve raid box office ocean twelve ...	1

Train Test Split

In [470]:
```
X_train, X_test, y_train, y_test = train_test_split(df['Content_Parsed'],
                                                    df['Category_Code'],
                                                    test_size=0.15,
                                                    random_state=8)
```

##Text Representation

In [471]:
```
# Parameter election
ngram_range = (1,2)
min_df = 10
max_df = 1.
max_features = 300
```

In [472]:
```
tfidf = TfidfVectorizer(encoding='utf-8',
                        ngram_range=ngram_range,
                        stop_words=None,
                        lowercase=False,
                        max_df=max_df,
                        min_df=min_df,
                        max_features=max_features,
                        norm='l2',
                        sublinear_tf=True)

features_train = tfidf.fit_transform(X_train).toarray()
labels_train = y_train
print(features_train.shape)

features_test = tfidf.transform(X_test).toarray()
labels_test = y_test
print(features_test.shape)
```

```
(1891, 300)
(334, 300)
```

In [473]:

```
for Product, category_id in sorted(category_codes.items()):
    features_chi2 = chi2(features_train, labels_train == category_id)
    indices = np.argsort(features_chi2[0])
    feature_names = np.array(tfidf.get_feature_names())[indices]
    unigrams = [v for v in feature_names if len(v.split(' ')) == 1]
    bigrams = [v for v in feature_names if len(v.split(' ')) == 2]
    print("# '{}' category:".format(Product))
    print("  . Most correlated unigrams:\n. {}".format('\n. '.join(unigrams[-
    print("  . Most correlated bigrams:\n. {}".format('\n. '.join(bigrams[-2:
    print("")
```

```
# 'business' category:
  . Most correlated unigrams:
. market
. price
. economy
. growth
. bank
  . Most correlated bigrams:
. last year
. year old

# 'entertainment' category:
  . Most correlated unigrams:
. tv
. music
. award
. star
. film
  . Most correlated bigrams:
. mr blair
. prime minister

# 'politics' category:
  . Most correlated unigrams:
. minister
. blair
. election
. party
. labour
  . Most correlated bigrams:
. prime minister
. mr blair

# 'sport' category:
  . Most correlated unigrams:
. game
. win
. team
. cup
. match
  . Most correlated bigrams:
. say mr
. year old

# 'tech' category:
```

```
  . Most correlated unigrams:
. digital
. computer
. technology
. software
. users
  . Most correlated bigrams:
. year old
. say mr
```

In [474]:
```
bigrams
```

Out[474]:
```
['tell bbc', 'last year', 'mr blair', 'prime minister', 'year old', 'say m
r']
```

In [475]:
```
# X_train
with open('NewPickle/X_train.pickle', 'wb') as output:
    pickle.dump(X_train, output)

# X_test
with open('NewPickle/X_test.pickle', 'wb') as output:
    pickle.dump(X_test, output)

# y_train
with open('NewPickle/y_train.pickle', 'wb') as output:
    pickle.dump(y_train, output)

# y_test
with open('NewPickle/y_test.pickle', 'wb') as output:
    pickle.dump(y_test, output)

# df
with open('NewPickle/df.pickle', 'wb') as output:
    pickle.dump(df, output)

# features_train
with open('NewPickle/features_train.pickle', 'wb') as output:
    pickle.dump(features_train, output)

# labels_train
with open('NewPickle/labels_train.pickle', 'wb') as output:
    pickle.dump(labels_train, output)

# features_test
with open('NewPickle/features_test.pickle', 'wb') as output:
    pickle.dump(features_test, output)

# labels_test
with open('NewPickle/labels_test.pickle', 'wb') as output:
    pickle.dump(labels_test, output)

# TF-IDF object
with open('NewPickles/tfidf.pickle', 'wb') as output:
    pickle.dump(tfidf, output)
```

#04

In [476]:

```
# Dataframe
path_df = "NewPickle/df.pickle"
with open(path_df, 'rb') as data:
    df = pickle.load(data)

# features_train
path_features_train = "NewPickle/features_train.pickle"
with open(path_features_train, 'rb') as data:
    features_train = pickle.load(data)

# labels_train
path_labels_train = "NewPickle/labels_train.pickle"
with open(path_labels_train, 'rb') as data:
    labels_train = pickle.load(data)

# features_test
path_features_test = "NewPickle/features_test.pickle"
with open(path_features_test, 'rb') as data:
    features_test = pickle.load(data)

# labels_test
path_labels_test = "NewPickle/labels_test.pickle"
with open(path_labels_test, 'rb') as data:
    labels_test = pickle.load(data)

print(features_train.shape)
print(features_test.shape)
```

```
(1891, 300)
(334, 300)
```

Random Forest

In [477]:

```
rf_0 = RandomForestClassifier(random_state = 8)

print('Parameters currently in use:\n')
pprint(rf_0.get_params())
```

```
Parameters currently in use:

{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 8,
 'verbose': 0,
 'warm_start': False}
```

In [478]:

```
# n_estimators
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 1000, num = 5

# max_features
max_features = ['auto', 'sqrt']

# max_depth
max_depth = [int(x) for x in np.linspace(20, 100, num = 5)]
max_depth.append(None)

# min_samples_split
min_samples_split = [2, 5, 10]

# min_samples_leaf
min_samples_leaf = [1, 2, 4]

# bootstrap
bootstrap = [True, False]

# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}

pprint(random_grid)
```

```
{'bootstrap': [True, False],
 'max_depth': [20, 40, 60, 80, 100, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [200, 400, 600, 800, 1000]}
```

In [480]:
```
# First create the base model to tune
rfc = RandomForestClassifier(random_state=8)

# Definition of the random search
random_search = RandomizedSearchCV(estimator=rfc,
                                   param_distributions=random_grid,
                                   n_iter=50,
                                   scoring='accuracy',
                                   cv=3,
                                   verbose=1,
                                   random_state=8)

# Fit the random search model
random_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 50 candidates, totalling 150 fits
```

Out[480]:
```
RandomizedSearchCV(cv=3, estimator=RandomForestClassifier(random_state=8),
                   n_iter=50,
                   param_distributions={'bootstrap': [True, False],
                                        'max_depth': [20, 40, 60, 80, 100,
                                                      None],
                                        'max_features': ['auto', 'sqrt'],
                                        'min_samples_leaf': [1, 2, 4],
                                        'min_samples_split': [2, 5, 10],
                                        'n_estimators': [200, 400, 600, 80
0,
                                                         1000]},
                   random_state=8, scoring='accuracy', verbose=1)
```

In [481]:
```
print("The best hypepara from Random Search are:")
print(random_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(random_search.best_score_)
```

```
The best hyperparameters from Random Search are:
{'n_estimators': 400, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_f
eatures': 'sqrt', 'max_depth': 40, 'bootstrap': False}

The mean accuracy of a model with these hyperparameters is:
0.9423632597959063
```

In [482]:

```
# Create the parameter grid based on the results of random search
bootstrap = [False]
max_depth = [30, 40, 50]
max_features = ['sqrt']
min_samples_leaf = [1, 2, 4]
min_samples_split = [5, 10, 15]
n_estimators = [800]

param_grid = {
    'bootstrap': bootstrap,
    'max_depth': max_depth,
    'max_features': max_features,
    'min_samples_leaf': min_samples_leaf,
    'min_samples_split': min_samples_split,
    'n_estimators': n_estimators
}

# Create a base model
rfc = RandomForestClassifier(random_state=8)

# Manually create the splits in CV in order to be able to fix a random_state
cv_sets = ShuffleSplit(n_splits = 3, test_size = .33, random_state = 8)

# Instantiate the grid search model
grid_search = GridSearchCV(estimator=rfc,
                           param_grid=param_grid,
                           scoring='accuracy',
                           cv=cv_sets,
                           verbose=1)

# Fit the grid search to the data
grid_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 27 candidates, totalling 81 fits
```

Out[482]:

```
GridSearchCV(cv=ShuffleSplit(n_splits=3, random_state=8, test_size=0.33, tr
ain_size=None),
             estimator=RandomForestClassifier(random_state=8),
             param_grid={'bootstrap': [False], 'max_depth': [30, 40, 50],
                         'max_features': ['sqrt'],
                         'min_samples_leaf': [1, 2, 4],
                         'min_samples_split': [5, 10, 15],
                         'n_estimators': [800]},
             scoring='accuracy', verbose=1)
```

In [483]:
```
print("The best hypepara from Grid Search are:")
print(grid_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(grid_search.best_score_)
```

```
The best hyperparameters from Grid Search are:
{'bootstrap': False, 'max_depth': 30, 'max_features': 'sqrt', 'min_samples_
leaf': 1, 'min_samples_split': 10, 'n_estimators': 800}

The mean accuracy of a model with these hyperparameters is:
0.9402666666666667
```

In [484]:
```
best_rfc = grid_search.best_estimator_
```

In [485]:
```
best_rfc
```

Out[485]:
```
RandomForestClassifier(bootstrap=False, max_depth=30, max_features='sqrt',
                       min_samples_split=10, n_estimators=800, random_state
=8)
```

In [486]:
```
best_rfc.fit(features_train, labels_train)
```

Out[486]:
```
RandomForestClassifier(bootstrap=False, max_depth=30, max_features='sqrt',
                       min_samples_split=10, n_estimators=800, random_state
=8)
```

In [ ]:
```
rfc_pred = best_rfc.predict(features_test)
```

In [487]:
```
# Training accuracy
print("The training accuracy is: ")
print(accuracy_score(labels_train, best_rfc.predict(features_train)))
```

```
The training accuracy is:
1.0
```

In [488]:
```
# Test accuracy
print("The test accuracy is: ")
print(accuracy_score(labels_test, rfc_pred))
```

```
The test accuracy is:
0.20958083832335328
```

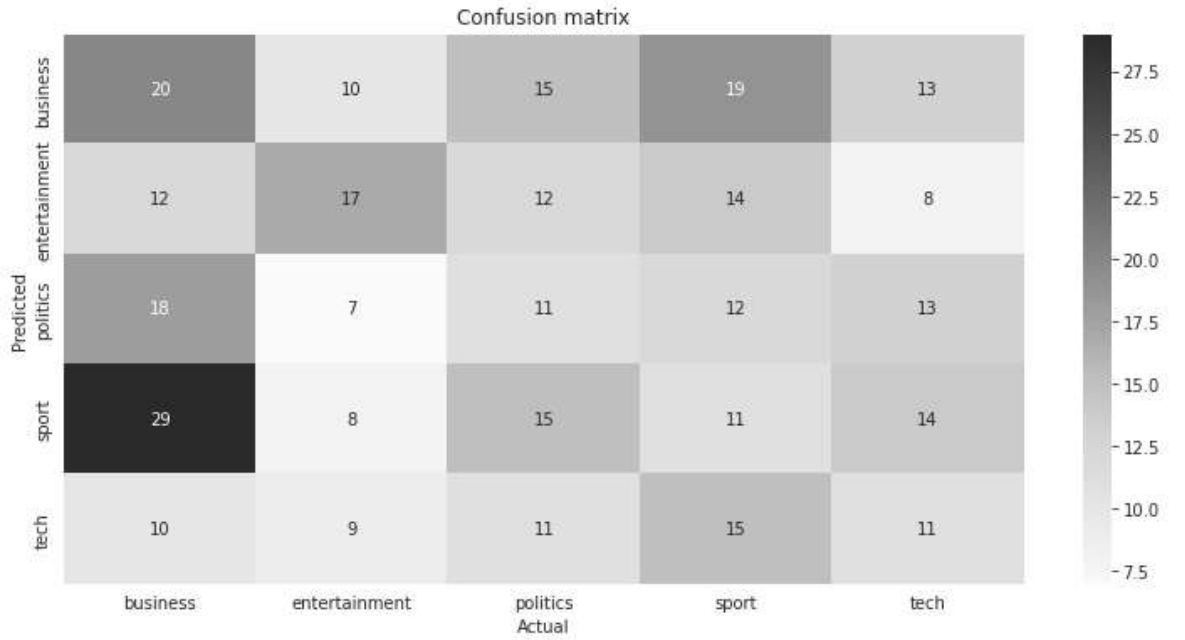
In [489]:

```
# Classification report
print("Classification report")
print(classification_report(labels_test,rfc_pred))
```

```
Classification report
              precision    recall  f1-score   support

           0       0.22      0.26      0.24        77
           1       0.33      0.27      0.30        63
           2       0.17      0.18      0.18        61
           3       0.15      0.14      0.15        77
           4       0.19      0.20      0.19        56

    accuracy                           0.21       334
   macro avg       0.21      0.21      0.21       334
weighted avg       0.21      0.21      0.21       334
```

In [490]:

```
aux_df = df[['category', 'Category_Code']].drop_duplicates().sort_values('Cat
conf_matrix = confusion_matrix(labels_test, rfc_pred)
plt.figure(figsize=(12.8,6))
sns.heatmap(conf_matrix,
            annot=True,
            xticklabels=aux_df['category'].values,
            yticklabels=aux_df['category'].values,
            cmap="Blues")
plt.ylabel('Predicted')
plt.xlabel('Actual')
plt.title('Confusion matrix')
plt.show()
```



In [491]:
```
base_model = RandomForestClassifier(random_state = 8)
base_model.fit(features_train, labels_train)
accuracy_score(labels_test, base_model.predict(features_test))
```

Out[491]: 0.9550898203592815

In [492]:
```
best_rfc.fit(features_train, labels_train)
accuracy_score(labels_test, best_rfc.predict(features_test))
```

Out[492]: 0.9610778443113772

In [493]:
```
d = {
     'Model': 'Random Forest',
     'Training Set Accuracy': accuracy_score(labels_train, best_rfc.predict(f
     'Test Set Accuracy': accuracy_score(labels_test, rfc_pred)
}

df_models_rfc = pd.DataFrame(d, index=[0])
```

In [494]:
```
df_models_rfc
```

Out[494]:

	Model	Training Set Accuracy	Test Set Accuracy
0	Random Forest	1.0	0.209581

In [495]:
```
with open('NewModels/best_rfc.pickle', 'wb') as output:
    pickle.dump(best_rfc, output)

with open('NewModels/df_models_rfc.pickle', 'wb') as output:
    pickle.dump(df_models_rfc, output)
```

##SVM

In [496]:
```
svc_0 =svm.SVC(random_state=8)

print('Parameters currently in use:\n')
pprint(svc_0.get_params())
```

```
Parameters currently in use:

{'C': 1.0,
 'break_ties': False,
 'cache_size': 200,
 'class_weight': None,
 'coef0': 0.0,
 'decision_function_shape': 'ovr',
 'degree': 3,
 'gamma': 'scale',
 'kernel': 'rbf',
 'max_iter': -1,
 'probability': False,
 'random_state': 8,
 'shrinking': True,
 'tol': 0.001,
 'verbose': False}
```

In [497]:
```
# C
C = [.0001, .001, .01]

# gamma
gamma = [.0001, .001, .01, .1, 1, 10, 100]

# degree
degree = [1, 2, 3, 4, 5]

# kernel
kernel = ['linear', 'rbf', 'poly']

# probability
probability = [True]

# Create the random grid
random_grid = {'C': C,
               'kernel': kernel,
               'gamma': gamma,
               'degree': degree,
               'probability': probability
              }

pprint(random_grid)
```

```
{'C': [0.0001, 0.001, 0.01],
 'degree': [1, 2, 3, 4, 5],
 'gamma': [0.0001, 0.001, 0.01, 0.1, 1, 10, 100],
 'kernel': ['linear', 'rbf', 'poly'],
 'probability': [True]}
```

In [498]:
```
# First create the base model to tune
svc = svm.SVC(random_state=8)

# Definition of the random search
random_search = RandomizedSearchCV(estimator=svc,
                                   param_distributions=random_grid,
                                   n_iter=50,
                                   scoring='accuracy',
                                   cv=3,
                                   verbose=1,
                                   random_state=8)

# Fit the random search model
random_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 50 candidates, totalling 150 fits
```

Out[498]:
```
RandomizedSearchCV(cv=3, estimator=SVC(random_state=8), n_iter=50,
                   param_distributions={'C': [0.0001, 0.001, 0.01],
                                        'degree': [1, 2, 3, 4, 5],
                                        'gamma': [0.0001, 0.001, 0.01, 0.1,
1,
                                                  10, 100],
                                        'kernel': ['linear', 'rbf', 'pol
y'],
                                        'probability': [True]},
                   random_state=8, scoring='accuracy', verbose=1)
```

In [499]:
```
print("The best hypepara from Random Search are:")
print(random_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(random_search.best_score_)
```

```
The best hyperparameters from Random Search are:
{'probability': True, 'kernel': 'poly', 'gamma': 10, 'degree': 4, 'C': 0.0
1}

The mean accuracy of a model with these hyperparameters is:
0.9217434323614989
```

In [500]:
```
# Create the parameter grid based on the results of random search
C = [.0001, .001, .01, .1]
degree = [3, 4, 5]
gamma = [1, 10, 100]
probability = [True]

param_grid = [
  {'C': C, 'kernel':['linear'], 'probability':probability},
  {'C': C, 'kernel':['poly'], 'degree':degree, 'probability':probability},
  {'C': C, 'kernel':['rbf'], 'gamma':gamma, 'probability':probability}
]

# Create a base model
svc = svm.SVC(random_state=8)

# Manually create the splits in CV in order to be able to fix a random_state
cv_sets = ShuffleSplit(n_splits = 3, test_size = .33, random_state = 8)

# Instantiate the grid search model
grid_search = GridSearchCV(estimator=svc,
                           param_grid=param_grid,
                           scoring='accuracy',
                           cv=cv_sets,
                           verbose=1)

# Fit the grid search to the data
grid_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 28 candidates, totalling 84 fits
```

Out[500]:
```
GridSearchCV(cv=ShuffleSplit(n_splits=3, random_state=8, test_size=0.33, tr
ain_size=None),
             estimator=SVC(random_state=8),
             param_grid=[{'C': [0.0001, 0.001, 0.01, 0.1], 'kernel': ['line
ar'],
                          'probability': [True]},
                         {'C': [0.0001, 0.001, 0.01, 0.1], 'degree': [3, 4,
5],
                          'kernel': ['poly'], 'probability': [True]},
                         {'C': [0.0001, 0.001, 0.01, 0.1],
                          'gamma': [1, 10, 100], 'kernel': ['rbf'],
                          'probability': [True]}],
             scoring='accuracy', verbose=1)
```

In [501]:
```
print("The best hypepara from Grid Search are:")
print(grid_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(grid_search.best_score_)
```

```
The best hyperparameters from Grid Search are:
{'C': 0.1, 'kernel': 'linear', 'probability': True}

The mean accuracy of a model with these hyperparameters is:
0.9413333333333332
```

In [502]:
```
best_svc = grid_search.best_estimator_
best_svc
```

Out[502]:
```
SVC(C=0.1, kernel='linear', probability=True, random_state=8)
```

In [503]:
```
best_svc.fit(features_train, labels_train)
```

Out[503]:
```
SVC(C=0.1, kernel='linear', probability=True, random_state=8)
```

In [504]:
```
svc_pred = best_svc.predict(features_test)
```

In [505]:
```
# Train accuracy
print("The training accuracy is: ")
print(accuracy_score(labels_train, best_svc.predict(features_train)))
```

```
The training accuracy is:
0.958223162347964
```

In [506]:
```
# Test accuracy
print("The test accuracy is: ")
print(accuracy_score(labels_test, svc_pred))
```

```
The test accuracy is:
0.9580838323353293
```

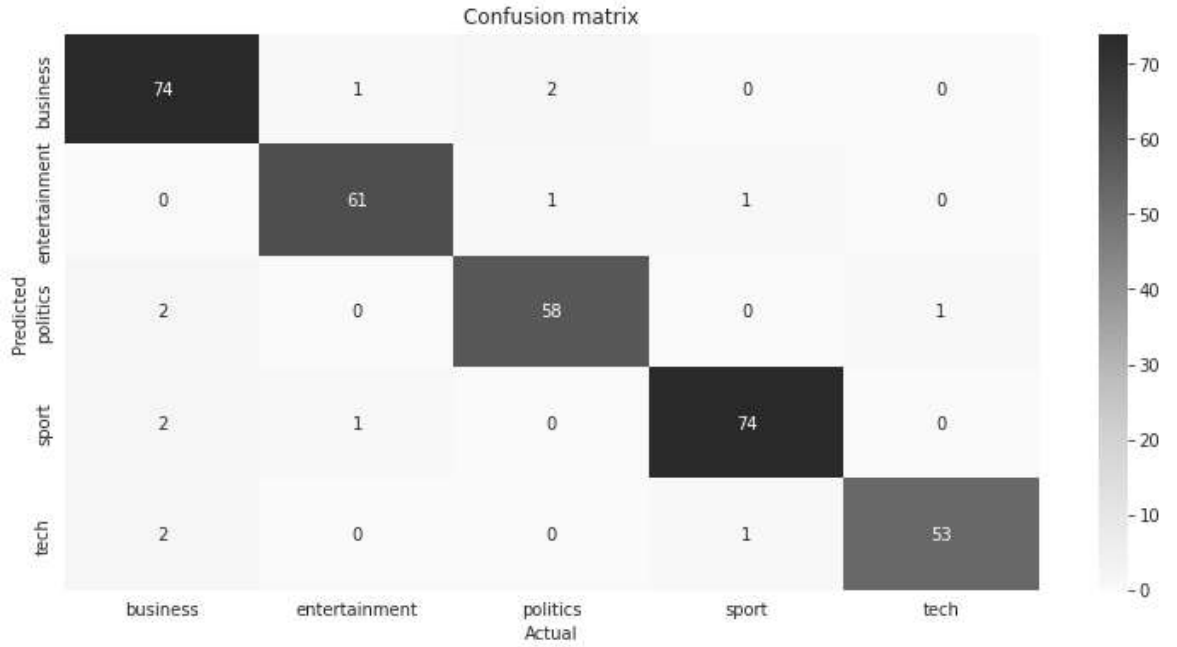
In [507]:

```
# Classification report
print("Classification report")
print(classification_report(labels_test,svc_pred))
```

```
Classification report
              precision    recall  f1-score   support

           0       0.93      0.96      0.94        77
           1       0.97      0.97      0.97        63
           2       0.95      0.95      0.95        61
           3       0.97      0.96      0.97        77
           4       0.98      0.95      0.96        56

    accuracy                           0.96       334
   macro avg       0.96      0.96      0.96       334
weighted avg       0.96      0.96      0.96       334
```

In [508]:

```
aux_df = df[['category', 'Category_Code']].drop_duplicates().sort_values('Cat
conf_matrix = confusion_matrix(labels_test, svc_pred)
plt.figure(figsize=(12.8,6))
sns.heatmap(conf_matrix,
            annot=True,
            xticklabels=aux_df['category'].values,
            yticklabels=aux_df['category'].values,
            cmap="Blues")
plt.ylabel('Predicted')
plt.xlabel('Actual')
plt.title('Confusion matrix')
plt.show()
```



In [509]:
```
base_model = svm.SVC(random_state = 8)
base_model.fit(features_train, labels_train)
accuracy_score(labels_test, base_model.predict(features_test))
```

Out[509]:
```
0.9730538922155688
```

In [510]:
```
best_svc.fit(features_train, labels_train)
accuracy_score(labels_test, best_svc.predict(features_test))
```

Out[510]:
```
0.9580838323353293
```

In [511]:
```
d = {
     'Model': 'SVM',
     'Training Set Accuracy': accuracy_score(labels_train, best_svc.predict(f
     'Test Set Accuracy': accuracy_score(labels_test, svc_pred)
}

df_models_svc = pd.DataFrame(d, index=[0])
df_models_svc
```

Out[511]:

	Model	Training Set Accuracy	Test Set Accuracy
0	SVM	0.958223	0.958084

In [512]:
```
with open('NewModels/best_svc.pickle', 'wb') as output:
    pickle.dump(best_svc, output)

with open('NewModels/df_models_svc.pickle', 'wb') as output:
    pickle.dump(df_models_svc, output)
```

##KNN

In [513]:
```
knnc_0 =KNeighborsClassifier()

print('Parameters currently in use:\n')
pprint(knnc_0.get_params())
```

```
Parameters currently in use:

{'algorithm': 'auto',
 'leaf_size': 30,
 'metric': 'minkowski',
 'metric_params': None,
 'n_jobs': None,
 'n_neighbors': 5,
 'p': 2,
 'weights': 'uniform'}
```

In [514]:
```
# Create the parameter grid
n_neighbors = [int(x) for x in np.linspace(start = 1, stop = 500, num = 100)]

param_grid = {'n_neighbors': n_neighbors}

# Create a base model
knnc = KNeighborsClassifier()

# Manually create the splits in CV in order to be able to fix a random_state
cv_sets = ShuffleSplit(n_splits = 3, test_size = .33, random_state = 8)

# Instantiate the grid search model
grid_search = GridSearchCV(estimator=knnc,
                           param_grid=param_grid,
                           scoring='accuracy',
                           cv=cv_sets,
                           verbose=1)

# Fit the grid search to the data
grid_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 100 candidates, totalling 300 fits
```

Out[514]:
```
GridSearchCV(cv=ShuffleSplit(n_splits=3, random_state=8, test_size=0.33, tr
ain_size=None),
             estimator=KNeighborsClassifier(),
             param_grid={'n_neighbors': [1, 6, 11, 16, 21, 26, 31, 36, 41,
46,
                                         51, 56, 61, 66, 71, 76, 81, 86, 9
1, 96,
                                         101, 106, 111, 116, 121, 127, 132,
137,
                                         142, 147, ...]},
             scoring='accuracy', verbose=1)
```

In [515]:
```
print("The best hypepara from Grid Search are:")
print(grid_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(grid_search.best_score_)
```

```
The best hyperparameters from Grid Search are:
{'n_neighbors': 11}

The mean accuracy of a model with these hyperparameters is:
0.9418666666666667
```

In [516]:
```
n_neighbors = [1,2,3,4,5,6,7,8,9,10,11]
param_grid = {'n_neighbors': n_neighbors}

knnc = KNeighborsClassifier()
cv_sets = ShuffleSplit(n_splits = 3, test_size = .33, random_state = 8)

grid_search = GridSearchCV(estimator=knnc,
                           param_grid=param_grid,
                           scoring='accuracy',
                           cv=cv_sets,
                           verbose=1)

grid_search.fit(features_train, labels_train)
```

```
Fitting 3 folds for each of 11 candidates, totalling 33 fits
```

Out[516]:
```
GridSearchCV(cv=ShuffleSplit(n_splits=3, random_state=8, test_size=0.33, tr
ain_size=None),
             estimator=KNeighborsClassifier(),
             param_grid={'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1
1]},
             scoring='accuracy', verbose=1)
```

In [517]:
```
print("The best hyperpara from Grid Search are:")
print(grid_search.best_params_)
print("")
print("The mean accuracy of a model with these hypepara is:")
print(grid_search.best_score_)
```

```
The best hyperparameters from Grid Search are:
{'n_neighbors': 10}

The mean accuracy of a model with these hyperparameters is:
0.9461333333333334
```

In [518]:
```
best_knnc = grid_search.best_estimator_
best_knnc
```

Out[518]:
```
KNeighborsClassifier(n_neighbors=10)
```

In [519]:
```
best_knnc.fit(features_train, labels_train)
```

Out[519]:
```
KNeighborsClassifier(n_neighbors=10)
```

In [520]:
```
knnc_pred = best_knnc.predict(features_test)
```

In [521]:
```
# Train accuracy
print("The training accuracy is: ")
print(accuracy_score(labels_train, best_knnc.predict(features_train)))
```

```
The training accuracy is:
0.952934955050238
```

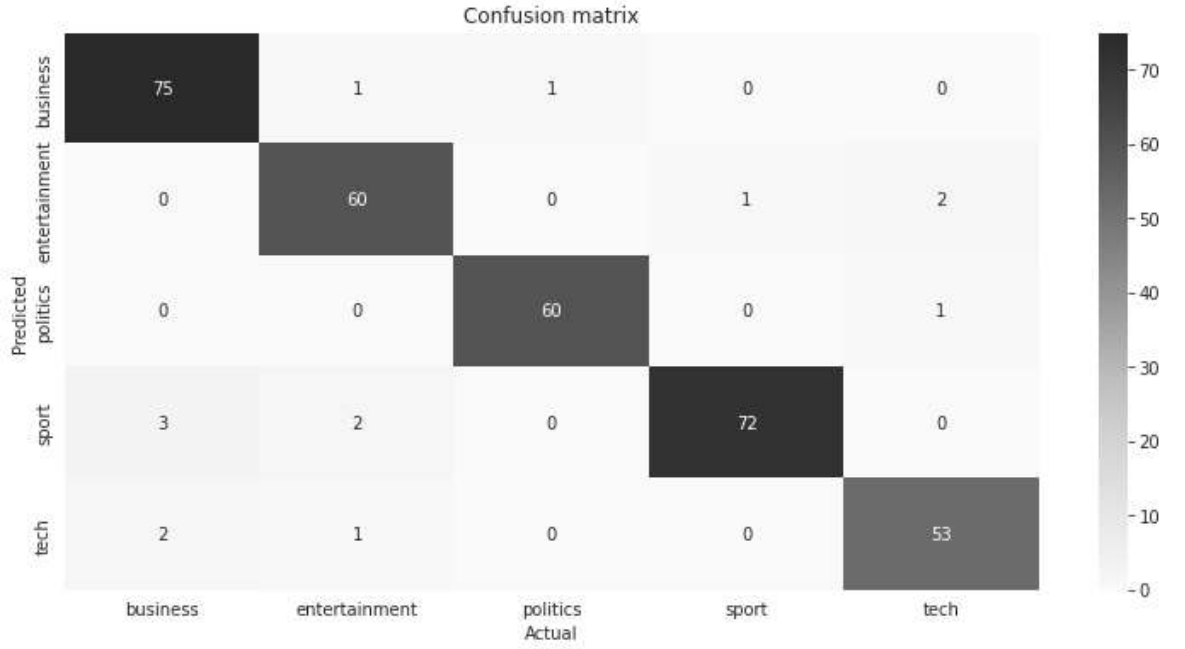
In [522]:
```
# Test accuracy
print("The test accuracy is: ")
print(accuracy_score(labels_test, knnc_pred))
```

```
The test accuracy is:
0.9580838323353293
```

In [523]:
```
# Classification report
print("Classification report")
print(classification_report(labels_test,knnc_pred))
```

```
Classification report
              precision    recall  f1-score   support

           0       0.94      0.97      0.96        77
           1       0.94      0.95      0.94        63
           2       0.98      0.98      0.98        61
           3       0.99      0.94      0.96        77
           4       0.95      0.95      0.95        56

    accuracy                           0.96       334
   macro avg       0.96      0.96      0.96       334
weighted avg       0.96      0.96      0.96       334
```

In [524]:
```
aux_df = df[['category', 'Category_Code']].drop_duplicates().sort_values('Cat
conf_matrix = confusion_matrix(labels_test, knnc_pred)
plt.figure(figsize=(12.8,6))
sns.heatmap(conf_matrix,
            annot=True,
            xticklabels=aux_df['category'].values,
            yticklabels=aux_df['category'].values,
            cmap="Blues")
plt.ylabel('Predicted')
plt.xlabel('Actual')
plt.title('Confusion matrix')
plt.show()
```



In [525]:
```
base_model = KNeighborsClassifier()
base_model.fit(features_train, labels_train)
accuracy_score(labels_test, base_model.predict(features_test))
```

Out[525]: 0.9491017964071856

In [526]:
```
best_knnc.fit(features_train, labels_train)
accuracy_score(labels_test, best_knnc.predict(features_test))
```

Out[526]: 0.9580838323353293

In [527]:
```
d = {
     'Model': 'KNN',
     'Training Set Accuracy': accuracy_score(labels_train, best_knnc.predict(
     'Test Set Accuracy': accuracy_score(labels_test, knnc_pred)
}

df_models_knnc = pd.DataFrame(d, index=[0])
```

In [528]:
```
df_models_knnc
```

Out[528]:

	Model	Training Set Accuracy	Test Set Accuracy
0	KNN	0.952935	0.958084

In [529]:
```
with open('NewModels/best_knnc.pickle', 'wb') as output:
    pickle.dump(best_knnc, output)

with open('NewModels/df_models_knnc.pickle', 'wb') as output:
    pickle.dump(df_models_knnc, output)
```

##Model Selection

In [530]:
```
path_pickles = "NewModels/"

list_pickles = [
    "df_models_knnc.pickle",
    "df_models_rfc.pickle",
    "df_models_svc.pickle"
]

df_summary = pd.DataFrame()

for pickle_ in list_pickles:

    path = path_pickles + pickle_

    with open(path, 'rb') as data:
        df = pickle.load(data)

    df_summary = df_summary.append(df)

df_summary = df_summary.reset_index().drop('index', axis=1)
df_summary
```

Out[530]:

	Model	Training Set Accuracy	Test Set Accuracy
0	KNN	0.952935	0.958084
1	Random Forest	1.000000	0.209581
2	SVM	0.958223	0.958084

In [531]:
```
df_summary.sort_values('Test Set Accuracy', ascending=False)
```

Out[531]:

	Model	Training Set Accuracy	Test Set Accuracy
0	KNN	0.952935	0.958084
2	SVM	0.958223	0.958084
1	Random Forest	1.000000	0.209581

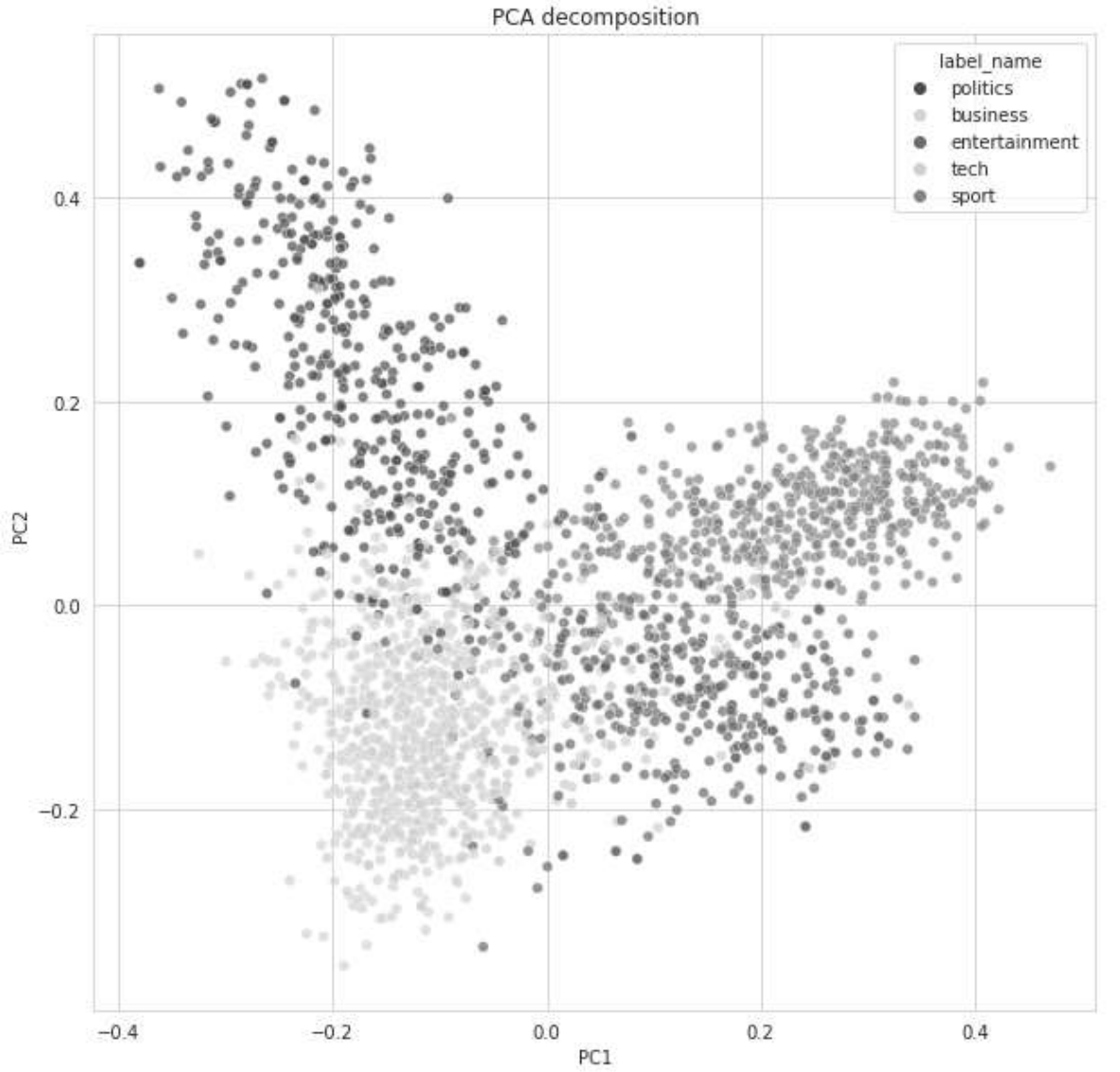
In [532]:
```
features = np.concatenate((features_train,features_test), axis=0)
labels = np.concatenate((labels_train,labels_test), axis=0)
print(features.shape)
print(labels.shape)
```

```
(2225, 300)
(2225,)
```

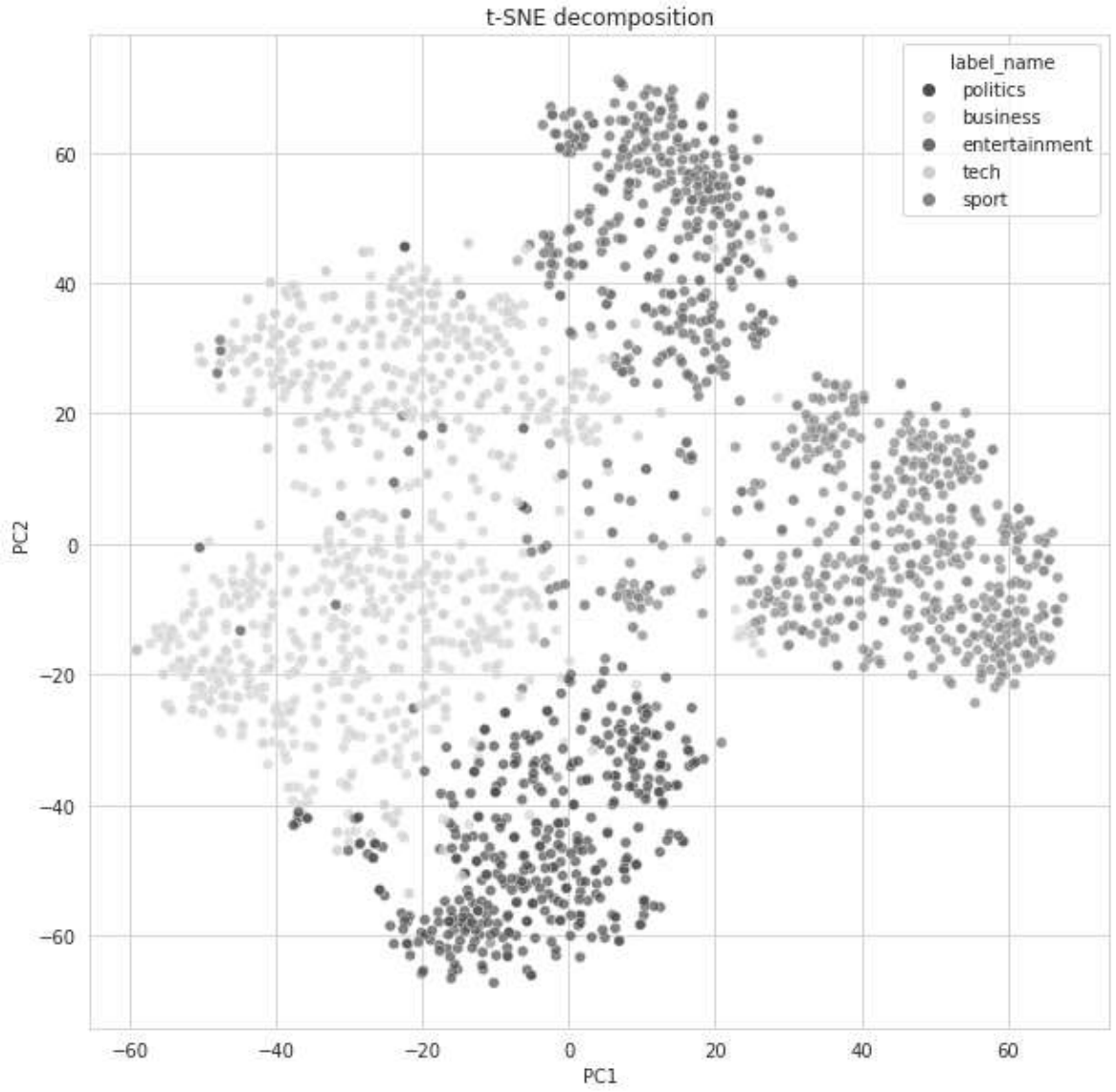
In [533]:

```
def plot_dim_red(model, features, labels, n_components=2):

    # Creation of the model
    if (model == 'PCA'):
        mod = PCA(n_components=n_components)
        title = "PCA decomposition"  # for the plot

    elif (model == 'TSNE'):
        mod = TSNE(n_components=2)
        title = "t-SNE decomposition"

    else:
        return "Error"

    # Fit and transform the features
    principal_components = mod.fit_transform(features)

    # Put them into a dataframe
    df_features = pd.DataFrame(data=principal_components,
                    columns=['PC1', 'PC2'])

    # Now we have to paste each row's label and its meaning
    # Convert labels array to df
    df_labels = pd.DataFrame(data=labels,
                             columns=['label'])

    df_full = pd.concat([df_features, df_labels], axis=1)
    df_full['label'] = df_full['label'].astype(str)

    # Get labels name
    category_names = {
        "0": 'business',
        "1": 'entertainment',
        "2": 'politics',
        "3": 'sport',
        "4": 'tech'
    }

    # And map labels
    df_full['label_name'] = df_full['label']
    df_full = df_full.replace({'label_name':category_names})

    # Plot
    plt.figure(figsize=(10,10))
    sns.scatterplot(x='PC1',
                    y='PC2',
                    hue="label_name",
                    data=df_full,
                    palette=["red", "pink", "royalblue", "greenyellow", "ligh
                    alpha=.7).set_title(title);
```

In [534]:
```
plot_dim_red("PCA",
             features=features,
             labels=labels,
             n_components=2)
```



In [535]:
```
plot_dim_red("TSNE",
             features=features,
             labels=labels,
             n_components=2)
```



In [536]:

```
# Dataframe
path_df = "NewPickles/df.pickle"
with open(path_df, 'rb') as data:
    df = pickle.load(data)

# SVM Model
path_model = "NewModels/best_knnc.pickle"
with open(path_model, 'rb') as data:
    knnc_model = pickle.load(data)

# Category mapping dictionary
category_codes = {
    'business': 0,
    'entertainment': 1,
    'politics': 2,
    'sport': 3,
    'tech': 4
}

category_names = {
    0: 'business',
    1: 'entertainment',
    2: 'politics',
    3: 'sport',
    4: 'tech'
}
```

In [537]:

```
predictions = knnc_model.predict(features_test)
# Indexes of the test set
index_X_test = X_test.index

print(index_X_test)


# We get them from the original df
df_test = df.loc[index_X_test]

# Add the predictions
df_test['Prediction'] = predictions

# Clean columns
df_test = df_test[['text', 'category', 'Category_Code', 'Prediction']]

# Decode
df_test['Category_Predicted'] = df_test['Prediction']
df_test = df_test.replace({'Category_Predicted':category_names})

# Clean columns again
df_test = df_test[['text', 'category', 'Category_Predicted']]
df_test.head()
```

```
Int64Index([1691, 1103,  477,  197,  475,  162,  887,  307, 1336, 1679,
            ...
            1567, 2130, 1216, 1135,  359,  393, 1746,  444, 2215,  733],
           dtype='int64', length=334)
```

Out[537]:

	text	category	Category_Predicted
1691	moya sidesteps davis cup in 2005 carlos moya h...	sport	sport
1103	poll idols face first hurdles vote for me - i...	politics	politics
477	britons fed up with net service a survey condu...	tech	tech
197	lib dems predict best ever poll the lib dems...	politics	politics
475	prince crowned top music earner prince earne...	entertainment	entertainment

In [538]:

```
condition = (df_test['category'] != df_test['Category_Predicted'])

df_misclassified = df_test[condition]

df_misclassified.head(3)
```

Out[538]:

	text	category	Category_Predicted
1144	mcdonald s to sponsor mtv show mcdonald s the...	business	entertainment
1565	ferdinand casts doubt over glazer rio ferdinan...	sport	business
535	pc ownership to double by 2010 the number of...	tech	business

In [539]:
```
def output_article(row_article):
    print('Actual Category: %s' %(row_article['category']))
    print('Predicted Category: %s' %(row_article['Category_Predicted']))
    print('----------------------------------------')
    print('Text: ')
    print('%s' %(row_article['text']))
```

In [540]:
```
random.seed(8)
list_samples = random.sample(list(df_misclassified.index), 3)
list_samples
```

Out[540]:
```
[1820, 1191, 1809]
```

In [541]:
```
output_article(df_misclassified.loc[list_samples[0]])
```

```
Actual Category: entertainment
Predicted Category: tech
----------------------------------------
Text:
johnny and denise lose passport johnny vaughan and denise van outen s satur
day night entertainment show passport to paradise will not return to screen
s  the bbc has said.  the ex-big breakfast presenters were recruited to hos
t the bbc one family variety show last july.  there are currently no plans
for another series   a spokeswoman said. she added the pair  brought a real
warmth to saturday night  but in the end we felt we had done enough with th
e format of the show .  passport to paradise involved a combination of game
s and outside broadcasts with a high level of audience participation. the f
irst instalment attracted more than 4.1 million viewers - but that had drop
ped to fewer than 2.7 million by the time it ended. the bbc spokeswoman sai
d graham norton s strictly dance fever would be a priority for 2005.  that
s very much on the cards for next year  and we re concentrating at the mome
nt on strictly come dancing  which is doing phenomenally well   she said.
```

In [542]:
```
output_article(df_misclassified.loc[list_samples[1]])
```

```
Actual Category: sport
Predicted Category: business
----------------------------------------
Text:
jones files lawsuit against conte marion jones has filed a lawsuit for defa
mation against balco boss victor conte following his allegations that he ga
ve her performance-enhancing drugs.  the sydney olympic gold medallist says
conte damaged her reputation and she is seeking $25m (£13m) in the suit. co
nte  whose company is at the centre of a doping investigation  made the cla
ims in a us television programme. he and three others were indicted in febr
uary by a federal grand jury for a variety of alleged offences. in an email
to the associated press on wednesday  conte said:  i stand by everything i
said . jones won three gold medals and two bronzes in sydney in 2000. her l
awsuit  filed in the us district court in san francisco  said the sprinter
had passed a lie detector test and that she  has never taken banned perform
ance-enhancing drugs . conte s statements  the suit added  were  false and
malicious . after the abc television program earlier this month  jones  law
yer richard nicholls said:  marion has steadfastly maintained her position
throughout: she has never  ever used performance-enhancing drugs.  victor c
onte is a man facing a 42-count federal indictment  while marion jones is o
ne of america s most decorated female athletes. mr conte s statements have
been wildly contradictory.  mr conte chose to make unsubstantiated allegati
ons on television  while marion jones demanded to take and then passed a li
e detector examination.  mr conte is simply not credible. we challenge him
to submit to the same lie detector procedure that marion jones passed.  the
sport s ruling body  the iaaf  is taking a cautious approach to conte s all
egations but contacted the us anti-doping agency. communications director n
ick davies said the iaaf would seek to contact conte  for further informati
on . but davies stressed it would be up to the american authorities to deci
de whether they will take action against jones in light of conte s televisi
on interview and the world governing body would monitor the situation close
ly.  if it is felt there is case to answer  it would be for its national go
verning body (usa track and field) to take the appropriate disciplinary act
ion  he added.  the us anti-doping agency has proved itself to be very dil
igent in its anti-doping war.  and i am sure  like ourselves  they will be
watching the television programme with great interest.  jones  who is under
investigation for steroid use by the us anti-doping agency  has continually
denied ever taking illegal substances since being investigated in the balco
scandal  although she praised a zinc supplement conte marketed. jones  who
did not win any medals in athens in august  has never failed a drugs test.
meanwhile  conte  who has been charged along with three other men of distri
buting illegal steroids and money laundering  is due to face trial in marc
h.
```

In [543]:
```
output_article(df_misclassified.loc[list_samples[2]])
```

```
Actual Category: business
Predicted Category: politics
----------------------------------------
Text:
golden rule  intact  says ex-aide chancellor gordon brown will meet his gol
den economic rule  with a margin to spare   according to his former chief e
conomic adviser.  formerly one of mr brown s closest treasury aides  ed bal
ls hinted at a budget giveaway on 16 march. he said he hoped more would be
done to build on current tax credit rules. any rate rise ahead of an expect
ed may election would not affect the labour party s chances of winning  he
added. last july  mr balls won the right to step down from his treasury pos
ition and run for parliament  defending the labour stronghold of normanton
in west yorkshire.  mr balls rejected the allegation that mr brown had been
sidelined in the election campaign  saying he was playing a  different  rol
e to the one he played in the last two elections. he rejected speculation t
hat mr brown was considering becoming foreign secretary  saying his recent
travels had been linked to efforts to boost international development. gord
on brown s decision to announce the date of the budget while on a trip to c
hina was a  sensible thing to do   since he was talking about skills and in
vestment at the time  mr balls told the bbc. commenting on speculation of a
n interest rate rise  he said it was not within the remit of the bank of en
gland s monetary policy committee (mpc) to factor a potential election into
its rate decisions. expectations of a rate rise have gathered pace after fi
gures showed that house prices are still rising. consumer borrowing rose at
a near-record pace in january.  i don t believe it would be a big election
issue in britain or a problem for labour   mr balls said. prime minister to
ny blair has yet to name the date of the election  but most pundits are bet
ting on 5 may as the likely day.
```

In [544]:
```
path_models = "NewModels/"

# SVM
path_svm = path_models + 'best_knnc.pickle'
with open(path_svm, 'rb') as data:
    knnc_model = pickle.load(data)

path_tfidf = "NewPickle/tfidf.pickle"
with open(path_tfidf, 'rb') as data:
    tfidf = pickle.load(data)

category_codes = {
    'business': 0,
    'entertainment': 1,
    'politics': 2,
    'sport': 3,
    'tech': 4
}
```

In [545]:
```
punctuation_signs = list("?:!.,;")
stop_words = list(stopwords.words('english'))

def create_features_from_text(text):

    # Dataframe creation
    lemmatized_text_list = []
    df = pd.DataFrame(columns=['text'])
    df.loc[0] = text
    df['Content_Parsed_1'] = df['text'].str.replace("\r", " ")
    df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace("\n", " ")
    df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace("    ", " ")
    df['Content_Parsed_1'] = df['Content_Parsed_1'].str.replace('"', '')
    df['Content_Parsed_2'] = df['Content_Parsed_1'].str.lower()
    df['Content_Parsed_3'] = df['Content_Parsed_2']
    for punct_sign in punctuation_signs:
        df['Content_Parsed_3'] = df['Content_Parsed_3'].str.replace(punct_sig
    df['Content_Parsed_4'] = df['Content_Parsed_3'].str.replace("'s", "")
    wordnet_lemmatizer = WordNetLemmatizer()
    lemmatized_list = []
    text = df.loc[0]['Content_Parsed_4']
    text_words = text.split(" ")
    for word in text_words:
        lemmatized_list.append(wordnet_lemmatizer.lemmatize(word, pos="v"))
    lemmatized_text = " ".join(lemmatized_list)
    lemmatized_text_list.append(lemmatized_text)
    df['Content_Parsed_5'] = lemmatized_text_list
    df['Content_Parsed_6'] = df['Content_Parsed_5']
    for stop_word in stop_words:
        regex_stopword = r"\b" + stop_word + r"\b"
        df['Content_Parsed_6'] = df['Content_Parsed_6'].str.replace(regex_stc
    df = df['Content_Parsed_6']
    df = df.rename({'Content_Parsed_6': 'Content_Parsed'})

    # TF-IDF
    features = tfidf.transform(df).toarray()

    return features
```

In [546]:
```
def get_category_name(category_id):
    for category, id_ in category_codes.items():
        if id_ == category_id:
            return category
```

In [547]:

```
def predict_from_text(text):

    # Predict using the input model
    prediction_svc = svc_model.predict(create_features_from_text(text))[0]
    prediction_svc_proba = svc_model.predict_proba(create_features_from_text(

    # Return result
    category_svc = get_category_name(prediction_svc)

    print("The predicted category using the SVM model is %s." %(category_svc)
    print("The conditional probability is: %a" %(prediction_svc_proba.max()*1
```

In [548]:

```
text = """ The center-right party Ciudadanos closed a deal on Wednesday with

The move would see the Socialist Party lose power in the region for the first

On Thursday, Marta Bosquet of Ciudadanos was voted in as the new speaker of t

The speaker’s role in the parliament is key for the calling of an investiture

Officially, the talks as to the make up of a future government have yet to st

The speaker’s role in the parliament is key for the calling of an investiture

The PP, which was ousted from power by the PSOE in the national government in

Wednesday was a day of intense talks among the parties in a bid to find a sol

The PSOE, meanwhile, argues that having won the elections with a seven-seat l


"""

predict_from_text(text)
```

```
The predicted category using the SVM model is business.
The conditional probability is: 55.09999028372485
```

In [549]:

```
# Politics

text = """Disputes have already broken out within the new political alliance

Just hours after the far-right Vox agreed to support the Popular Party (PP)'s

These early clashes suggest it could be difficult to export the model to othe

The PP and the liberal Ciudadanos have reached their own governing agreement

Ciudadanos has refused point-blank to meet with Vox representatives, but the

On Friday morning, Juan Marín of Ciudadanos said that there are no plans for

The reform party has insisted that the Vox-PP deal does not affect them at al

 Vox national leader Santiago Abascal (c) and Andalusian leader Francisco Ser
Vox national leader Santiago Abascal (c) and Andalusian leader Francisco Serr
But Vox insists on a family department, and said it will expect loyalty from

These early clashes suggest it could be difficult to export the model to othe

The PP is anxious to win back power in regions like Valencia, the Balearic Is

Parliamentary debate
The PSOE has already digested the fact that it is losing its hold on Spain's

The Socialists will not be putting forward a candidate, now that the PP nomin

The sum of the PP, Ciudadanos and Vox votes is four above the 55 required for

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is politics.
The conditional probability is: 66.28189176348307
```

In [550]:

```
in The New York Times’ list of 52 Places to Go in 2019. The recognition comes

 Cádiz for The New York Times’ list, lives in Spain himself and is no stranger

 a major maritime link between America and Europe, it’s not very well known to


aría.
ría.
trobar Saja River, recently opened on Santa Elena street, and Código de Barra

 of its own, including Restaurante Café Royalty, which opened opened in 1912,


iz).ampliar foto
z). NEIL FARRIN GETTY IMAGES
 de la Frontera, known for the fortified wines known in English as sherry. Wir




ry art sits between Barbate and Vejer de la Frontera. It is a private gallery

s with a few of its own:




rom February 28 to March 10. In fact it is so unique that it is applying to be



GETTY
features an old Roman theater, the old cathedral and stone arches that lead to



L GETTY
ith its ficus and palm trees, to the Provincial Museum containing Phoenician s


mpliar foto
EN WELSH GETTY
ts, the province of Cádiz has a long and action-packed history, while its capi


ameda.ampliar foto
meda. JUAN CARLOS TORO
ch as well as for its wineries, this coastal town has been described by journa


o
IAN GETTY
```

```
e it a perfect haunt for surfers of various descriptions. In less than an hour


marón de la Isla, named after the famous singer, has shows every week and is a


to
ES
long Cádiz’s Atlantic coast – La Breña, Los Alcornocales and el Estrecho – as


 in Tarifa, wind and kitesurfers can skid across the water with a view of Afri


e you through a string of white villages – Alcalá del Valle, Algar, Algodonale
```

```
The predicted category using the SVM model is politics.
The conditional probability is: 51.108910898771256
```

In [551]:

```
# Business

text = """
Vodafone España has informed representatives of its employees that it is putt

“In the current market climate, demand for services continues to grow exponen

Vodafone added that the current expectations of clients, “who demand an agile

As such, the company continued, it is looking to “reverse the negative trend

The operator says that it is sure it can reach a deal with labor unions so th

Vodafone has suffered a great deal in the trade war that was sparked by its r

In the first three quarters of 2018, Vodafone has lost 361,000 cellphone line

The operator executed a similar collective dismissal plan (known in Spanish a

Before the acquisition of ONO, Vodafone also executed an ERE in 2013. On that

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is business.
The conditional probability is: 47.16858388488912
```

In [552]:

```
# Tech

text = """
Elon Musk told the world in late 2017 that Tesla was taking its automotive kn

The German automaker also committed to manufacturing the truck this summer, w


While there are a few Tesla Semi prototypes on the road now, and a dozen or s

DAIMLER FIRST SHOWED OFF A PROTOTYPE IN 2015
This has left the door wide open for companies like Daimler, the parent compa

The new Cascadia is not much more advanced than the prototype was in 2015. In


The Freightliner Inspiration Truck at the event in 2015.
But the new Cascadia is doing this with a limited set of sensors. There's a f

This helps keep costs down, but means the technology is more in line with wha

DAIMLER'S TRUCK HAS MORE IN COMMON WITH NISSAN'S PROPILOT SYSTEM THAN TESLA'S
Keeping with a theme of less is more, there's also no camera-based monitoring

A sensor in the steering column measures resistance applied to the steering w



The new Cascadia is a far cry from a fully autonomous truck, but based on my

A Daimler representative also told me that, while lane centering is on, the d

RELATED

This is what it's like to ride in Daimler's self-driving semi truck
Daimler promised some other modern technologies are coming the new Cascadia,

The Cascadia won't be as stuffed with tech as the Tesla Semi, nor is it as sl
"""

predict_from_text(text)
```

```
The predicted category using the SVM model is business.
The conditional probability is: 67.73745032912375
```

In [553]:

```
# Sports

text = """
Spain has agreed to host the soccer final of the Copa Libertadores between Ar

The final in Madrid is a punch in the soul to all fans of soccer in Argentina

ONLINE SPORTS DAILY OLE

The final was set to take place in Argentina but was suspended twice after fa

In view of the insecurity, the South American Football Confederation (Conmebo

 Embedded video

Sebastián Lisiecki
@sebalisiecki
 Así fue la llegada de Boca al Monumental. Pésimo la seguridad q los mete ent

575
7:23 PM - Nov 24, 2018
637 people are talking about this
Twitter Ads info and privacy
This was how Boca arrived at Monumental stadium. The security that got betwee

This is the first time a Copa Libertadores game has been played outside the A

But the feeling in Argentina has been less optimistic. The national newspaper

Security risk
In a message on Twitter, Sánchez promised that "security forces have extensiv

River and Boca have a long-standing rivalry fueled largely by the class divid

Scheduling issues
The final will take place on Sunday, December 9, on the final day of a three-

 Conmebol president Alejandro Domínguez on Tuesday.
Conmebol president Alejandro Domínguez on Tuesday.
Many details about the game have yet to be revealed, including how tickets wi

Conmebol and soccer club representatives began considering destinations for t



"""

predict_from_text(text)
```

```
The predicted category using the SVM model is tech.
The conditional probability is: 59.62510649552267
```

In [554]:

```
# Weather

text = """
A polar air mass that entered the Iberian peninsula on Wednesday has already

“An episode of intense cold” is forecast for Friday, when the mercury will co

Elsewhere, weather stations have recorded -8.2ºC in La Molina (Girona), at an

 Almería has rolled out vehicles to deal with wintry road conditions.
Almería has rolled out vehicles to deal with wintry road conditions. DIPUTACI
Aemet spokesman Rubén del Campo said that the cold spell is not out of the or

Temperatures have already dipped between six and eight degrees in a matter of

Temperatures on Friday and Saturday will be “very cold, with lows of five to

No snow
However, little to no snow is expected “not for lack of cold, but for lack of

Alerts are in place in Almería, Granada, Jaén, Aragón, Cantabria, Castilla-La

On Saturday, the orange warnings will extend to Córdoba, Salamanca, Valladoli


"""

predict_from_text(text)
```

```
The predicted category using the SVM model is entertainment.
The conditional probability is: 71.28419601310195
```

In [555]:

```
# Health

text = """
The obesity epidemic has been on the rise for years, with cases nearly tripli

An investigation by the Mar de Barcelona hospital has found that 80% of men a

Being overweight can mean a higher risk of suffering a number of diseases, in

The study, published in the Spanish Cardiology Magazine, points out that this

The issue, the experts state, is not an esthetic one, but rather a question o

Researchers at the Barcelona hospital revised all of the scientific literatur

There are currently 25 million people with excess weight, three million more

DR ALBERT GODAY, AUTHOR OF THE STUDY

“There are currently 25 million people with excess weight, three million more

“In men, excess weight is more usual up to the age of 50,” explains Goday. “F

The experts argue that any weight loss, no matter how small, reduces the risk

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is business.
The conditional probability is: 78.04540552025212
```

In [556]:

```
# Animal abuse

text = """
Spain’s animal rights party PACMA posted a 38-second video on Twitter on Frid

“Hunters shut what appears to be a fox in a cage and let it out only to peppe

 Video insertado

PACMA
✓
@PartidoPACMA
 Cazadores enjaulan a lo que parece ser un zorro y lo liberan solo para acrib

En realidad, son peligrosos psicópatas con escopeta y permiso de amas. #YoNoD

4.188
10:43 - 4 ene. 2019
7.443 personas están hablando de esto
Información y privacidad de Twitter Ads
At the start of the video, a man teases the caged animal with a stick. When t

The release of the video, which has had 255,000 views, coincided with the lau

As it notes on its website, PACMA is the only political group that opposes hu

No animal should die under fire. We will fight tirelessly until hunting becom

PACMA

The animal rights group is preparing a report to send to the regional governm

Last month, a Spanish hunter who was filmed while he chased and tortured a fo

And in November, animal rights groups and political parties reacted with indi

"""

predict_from_text(text)
```

```
The predicted category using the SVM model is business.
The conditional probability is: 64.06761565559422
```

In [ ]: