Exploratory Data Analysis

In this notebook we'll do some exploratory data analysis over our dataset. However, since we don't have our features created yet, we cannot do much at this point. In addition, when we create them, we won't be able to extract many insights because of the nature of text-based features. For this reason, only a shallow analysis will be done at this point.

For the plots we have used `seaborn` and `altair`. `altair` is a package which allows us to plot graphics with a simple grammar as we would do in ggplot2 or Tableau. It also provides easy-to-make interactive plots. For further information please visit the project site: [https://altair-viz.github.io/ (https://altair-viz.github.io/)](https://altair-viz.github.io/).

To install it, please type this command in the shell:

```
! conda install -c conda-forge altair vega_datasets notebook vega
```

In [1]:
```
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")
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-1-a60fa0867c28> in <module>
      4 import seaborn as sns
      5 sns.set_style("whitegrid")
----> 6 import altair as alt
      7 alt.renderers.enable("notebook")
      8

ModuleNotFoundError: No module named 'altair'
```

Loading the dataset:

In [ ]:
```
df_path = "/home/lnc/0. Latest News Classifier/01. Dataset Creation/"
df_path2 = df_path + 'News_dataset.csv'
df = pd.read_csv(df_path2, sep=';')
```

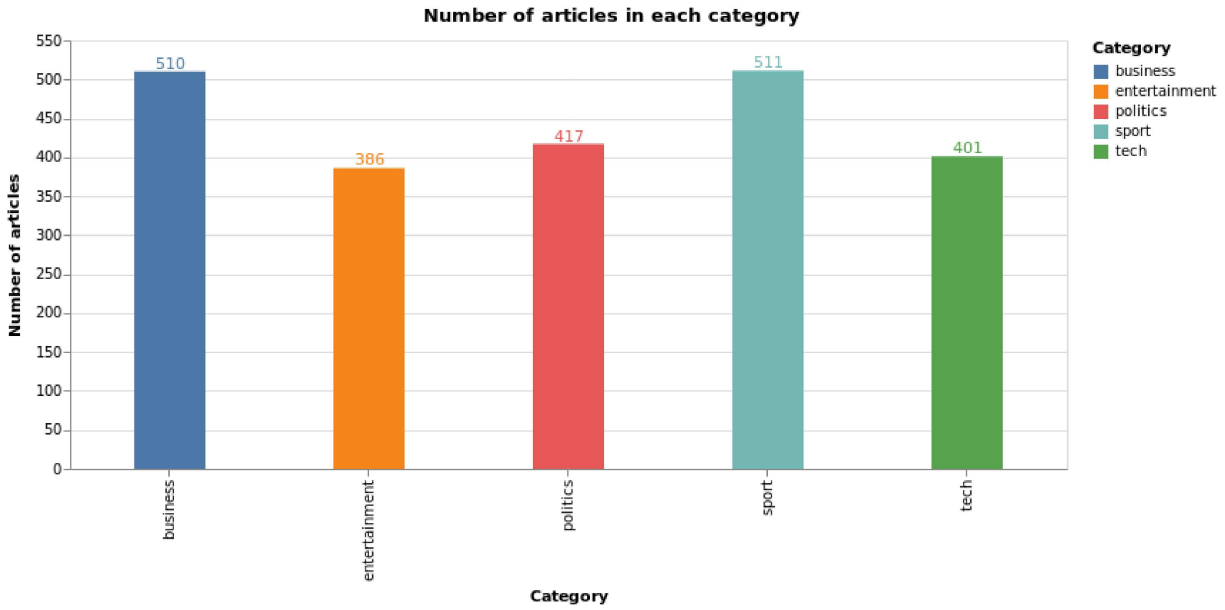
In [ ]:
```
df.head()
```

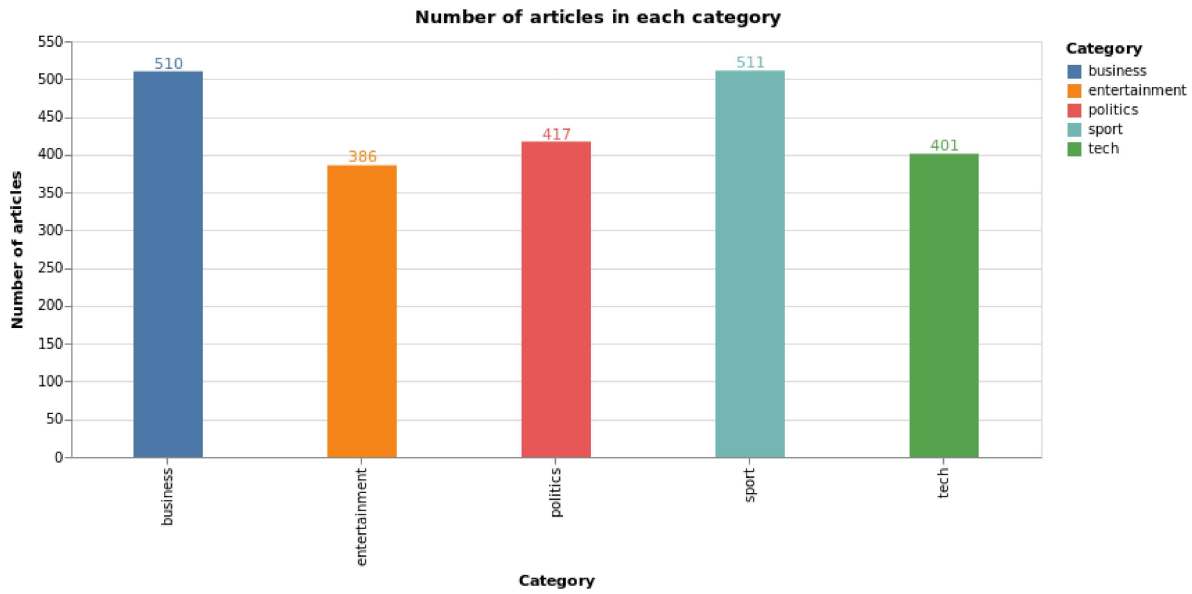
Number of articles in each category

In [4]:
```
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 0x7ffa5b9707f0>
```

Out[4]:



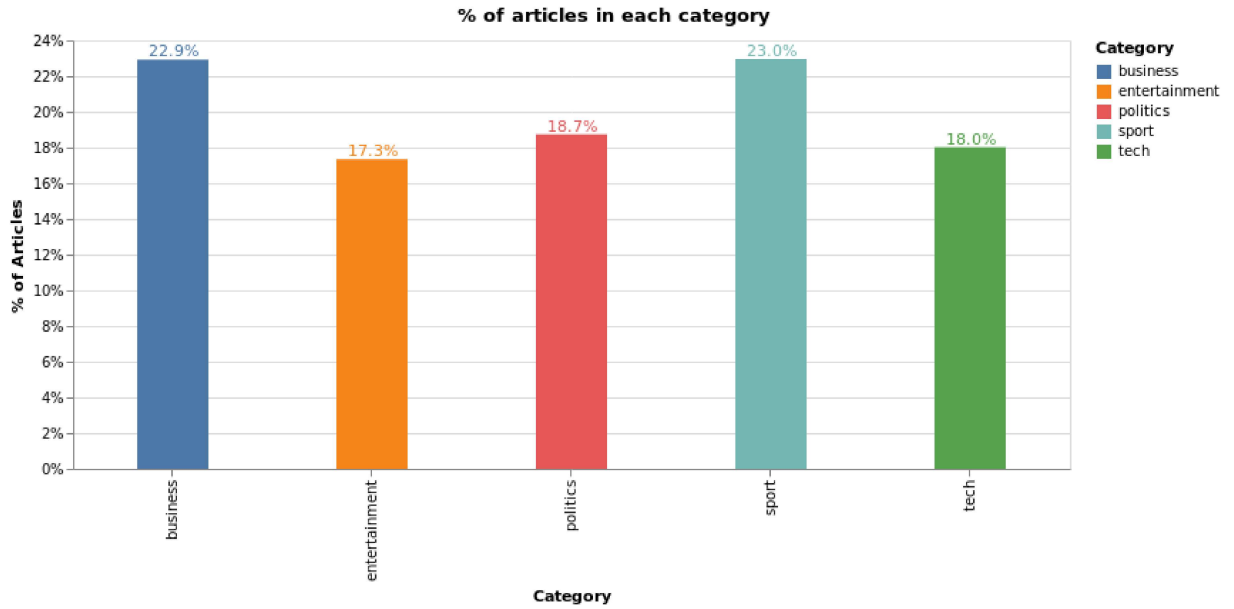


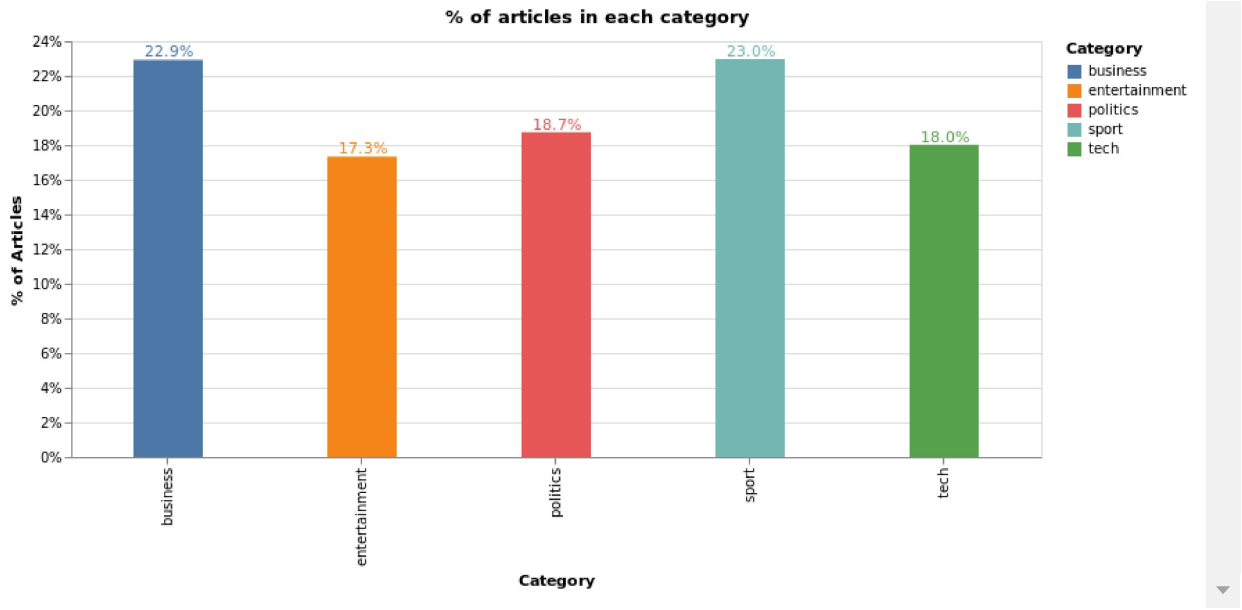
% of articles in each category

In [5]:
```
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 Articles'
    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 0x7ffa5b9707b8>
```

Out[5]:





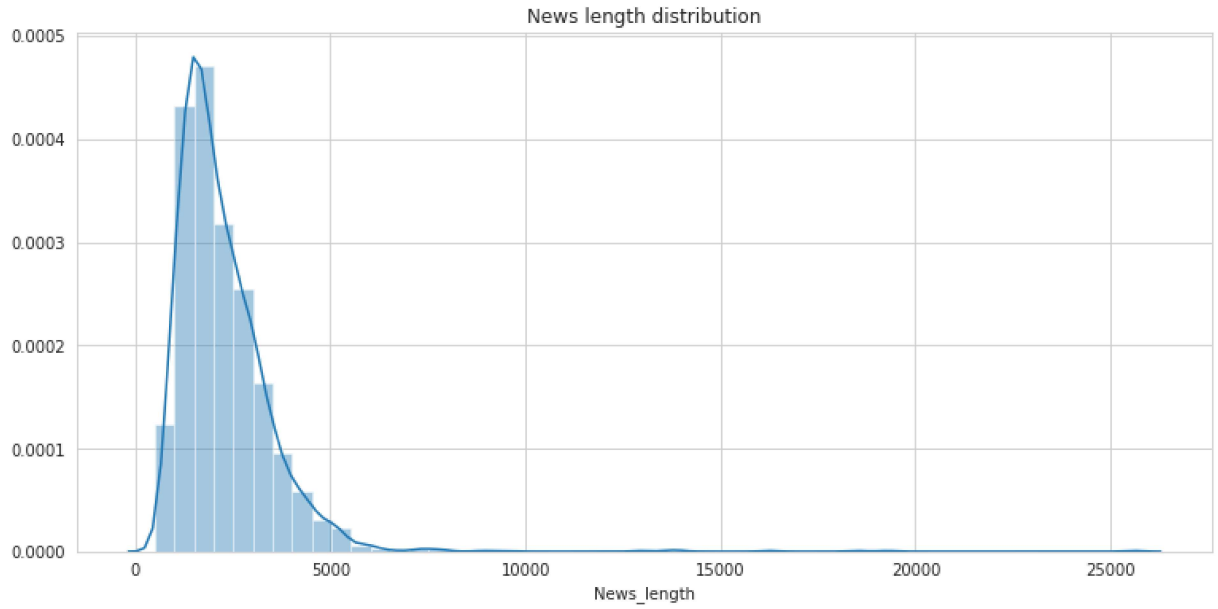
The classes are approximately balanced. We'll first try to train the models without oversampling/undersampling. If we see some bias in the model, we'll use these techniques.

News length by category

Definition of news length field. Although there are special characters in the text (`\r, \n`), it will be useful as an approximation.

In [6]:
```
df['News_length'] = df['Content'].str.len()
```

In [7]:
```
plt.figure(figsize=(12.8,6))
sns.distplot(df['News_length']).set_title('News length distribution');
```



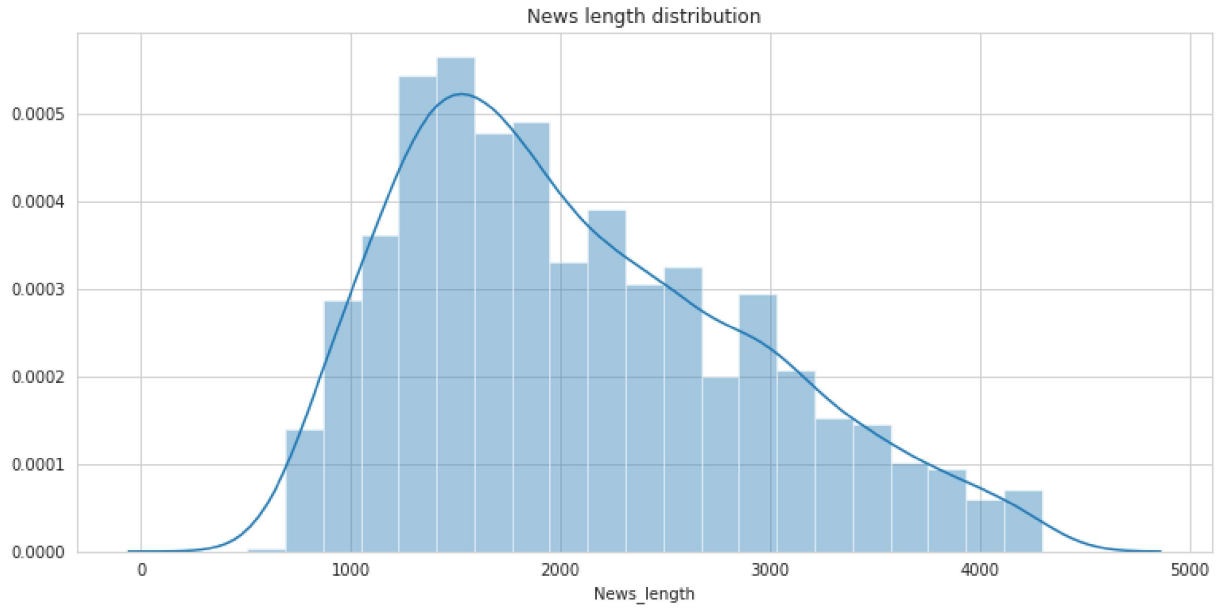
In [8]:
```
df['News_length'].describe()
```

Out[8]:
```
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
```

Let's remove from the 95% percentile onwards to better appreciate the histogram:

In [9]:
```
quantile_95 = df['News_length'].quantile(0.95)
df_95 = df[df['News_length'] < quantile_95]
```

In [10]:
```
plt.figure(figsize=(12.8,6))
sns.distplot(df_95['News_length']).set_title('News length distribution');
```



We can get the number of news articles with more than 10,000 characters:

In [11]:
```
df_more10k = df[df['News_length'] > 10000]
len(df_more10k)
```

Out[11]:
```
7
```

Let's see one:

In [12]:
```
df_more10k['Content'].iloc[0]
```

Out[12]:
```
'Scissor Sisters triumph at Brits\r\n\r\nUS band Scissor Sisters led the winner
s at the UK music industry\'s Brit Awards, walking off with three prizes. The f
lamboyant act scored a hat-trick in the international categories, winning the b
est group, best album and best newcomer awards. Glasgow group Franz Ferdinand w
on two prizes, as did Keane and Joss Stone, who was voted best urban act by dig
ital TV viewers. Robbie Williams\' Angels was named the best song of the past 2
5 years. Scissor Sisters frontwoman Ana Matronic collected the best internation
al album prize from singer Siouxsie Sioux. She told the audience: "If you told
us a year ago we would be getting these awards today we would have called you c
razy. You guys made our dream come true."\r\n\r\nThe band - whose self-titled L
P was 2004\'s biggest-selling album - thanked "all the members of the sisterhoo
d", adding: "We wouldn\'t be here without you." The US band, who opened the sho
w with Take Your Mama, won the best international act and newcomer awards, as w
ell as best international album.\r\n\r\nFranz Ferdinand, who were shortlisted i
n five categories, won best rock act and best British group, an award they dedi
cated to late DJ John Peel. But they missed out on best British live act, which
went to Muse. Keane won best British album and breakthrough act. Will Young won
the best single prize for Your Game. McFly won the best pop act prize, and Gwen
Stefani picked up the best international female artist award. Eminem won the ma
le prize.\r\n\r\nBest British male artist winner Mike Skinner - aka The Streets
- does not usually attend award ceremonies, but the Birmingham hip-hop artist p
erformed his hit Dry Your Eyes at the ceremony. However, he did not collect his
prize. A bandmate informed the crowd Skinner was "in the toilet". After beating
Amy Winehouse, Jamelia, Natasha Bedingfield and PJ Harvey to the best British f
emale prize, Joss Stone said: "I don\'t know what to say. I don\'t like doing t
his at all. I\'d like to thank my family for being really supportive and everyb
ody that made my record with me." "I don\'t even know what to do right now. Tha
nk you all you guys for voting for me, I feel sick right now." Viewers of digit
al music TV channel MTV Base voted Stone the winner in the best urban act categ
ory.\r\n\r\nLittle Britain comedy duo Matt Lucas and David Walliams presented t
he best song prize to Robbie Williams dressed as his former Take That colleague
s Gary Barlow and Howard Donald, leading him to quip he was "always the talente
d man of the band".\r\n\r\nWilliams\' track beat songs by Will Young, Queen, Ka
te Bush and Joy Division in a vote by BBC Radio 2 listeners to mark 25 years of
the UK music industry ceremony. It is his 15th Brit award, having already recei
ved 10 solo awards and four with Take That. He told the audience: "I\'m just am
azed that my career keeps going." Keane frontman Tom Chaplin thanked fans for e
nduring "rubbish gigs" after they won the British breakthrough act prize. He ad
ded: "A lot of people don\'t think it\'s cool that we\'ve had the guts to be ou
rselves but it\'s a vital part of who we are as a band and receiving this is re
cognition of that."\r\n\r\nNatasha Bedingfield - in the running for best Britis
h female and best pop act - performed with her brother Daniel for the first tim
e at Wednesday\'s event.\r\n\r\nThe chart-topping siblings duetted on the Chaka
Khan hit Ain\'t Nobody. Meanwhile, Joss Stone performed Right To Be Wrong backe
d by a gospel choir, while Lemar and Jamelia performed the Robert Palmer track
Addicted To Love. Bob Geldof won a prize for his outstanding contribution to mu
sic. Of the 15 Brit awards for achievements in 2004, 10 were won by artists tip
ped in the BBC News website\'s Sound of 2004 list of artists to watch, publishe
d at the start of last year. Scissor Sisters, Franz Ferdinand, Keane, Joss Ston
e and McFly were all in the Sound of 2004 top 10. The other five Brits winners
were already established before Sound of 2004 was compiled. The ceremony will b
e televised on ITV1 on Thursday.\r\n\r\n\r\n\r\nI\'m speechless. Best song of t
he last 25 years? Yeah right.\r\n\r\nI very much doubt that \'Angels\' was even
the best song of the week that it came out. Like every track Robbie has release
d as a single, it\'s a blatant but poor facsimile of something that someone els
```

e has done better before.\r\n\r\nGive us a break...!!!\r\n\r\nBest song in 25 y
ears, you must be joking. Its good if you like that sort of thing, but really!
\r\n\r\nListened to Angels on Radio 1 this morning when I was driving into wor
k. Had not heard it for a while. I love Robs voice, the lyrics and tune. Perfec
tion!\r\n\r\nAs usual, the public have short memories when it comes to voting f
or "the greatest". There must be more than a dozen songs in the last 25 years t
hat deserve this award more. It\'s not exactly groundbreaking. Presumably, the
age range that could be bothered to vote is is pretty low...\r\n\r\nI\'m actual
ly embarrassed to be British if that is the best song we have produced in the l
ast 25 years!!\r\n\r\nWhat about The Specials - Ghost Town, The Buzzcocks - Eve
r Fallen in Love With Someone... Happy Mondays - Kinky Afro, McAlmont & Butler
- Yes, Joy Division - Love Will Tear Us Apart... Angels is middle-of-the-road r
ubbish.\r\n\r\nAngels is a awful piece of sentimental claptrap. It\'s musically
and lyrically inept; and fantastically overrated, a bit like Mr Williams himsel
f. This result isn\'t very surprising though, The Brits has a long history of c
elebrating rubbish music!\r\n\r\nBest of the last 25 years? Maybe. Cunning to m
ake the timescale not include Stairway to Heaven or Bohemian Rhapsody, but it d
oes kind of make it a bit of a hollow award really. Not much competition in the
last 25 years after all.\r\n\r\nIt\'s alright for a pop song - but the best son
g of the last 25 years???\r\n\r\nThere is no way on earth that song should have
been voted the best of the last 25 years....it\'s a travesty.\r\n\r\nRubbish! W
ho voted it for it to be included in any list? I am a regular listener to Radio
2 but I don\'t recall the invitation to vote for this bland, slushy rubbish whi
ch might appeal to the masses who wouldn\'t know a good song if it jumped up an
d bit them on the nose but is certainly NOT the best song of the last 25 years.
How depressing and just when we thought manufactured \'pop\' was on the way out
- where on earth did this dreadful list appear from?\r\n\r\nWhile I am biassed
in that I thought Love Will Tear Us Apart should have won, in all seriousness,
I think that the best song of the last 25 years should not include songs less t
han 5 years old as that would exclude songs which are popular because of novelt
y. Then again, well done Robbie, good show.\r\n\r\nYou\'ve got to be kidding.\r
\n\r\nAngels is a great song, but not the best song of the last 25 years. Only
the best song to be up for nomination at the Brits.\r\n\r\nI think Angels is a
great song and deserved to be in the run up for this award but I don\'t think i
ts the best song from the past 25 years! Right enough, it is better than some o
f the others in this catergory, for example, what was Will Young doing being no
minated in the first place - he is alright but the song isn\'t that good! I\'m
happy for Robbie himself though!\r\n\r\nBest song in the last 25 years? What a
Joke! Think of all the great rock and pop songs released in the 80s and pretty
much all of them are better than Angels. Phil Collins doesn\'t deserve awards f
or all the good songs he wrote? Angels is an overrated song, that got tiresome
even before you had finished listening to it. Soppy rubbish at best. Hopefully
manufactured rubbish will die down soon, and let the real artists who worked ha
rd for there glory receive awards.\r\n\r\nSo boringly obvious and typical of th
e bland nature of mainstream music in Britain today, for me it\'s proof that mu
sic and democracy just don\'t mix. Still, at least it wasn\'t Will Young...\r\n
\r\nOh it\'s all just a bit of fun. People take these awards too seriously! Rob
bie has millions of loyal fans, while even non-fans know the words to angels. H
im winning obviously reflects who votes in these awards. Personally I wanted Wi
ll Young to win, but that was not really due to his musical talent!\r\n\r\nI ha
te the song, all it brings back is memories of school discos and no-one to slow
-dance with!\r\n\r\nI agree about Angels. I never get fed up hearing it. Whenev
er the song comes on the radio I turn the radio up, smile and sing along (very
badly, that is why the radio has to be turned up to drown my voice out). The so
ng makes me calm and serene and happy. Well done Robbie.\r\n\r\nI think that al
though Robbie Williams is a good performer, that Angels isn\'t really that good
a song. It certainly isn\'t anywhere near as good as Love Will Tear Us Apart by

Joy Division or Wuthering Heights by Kate Bush.\r\n\r\nAngels is a fantastic so ng. All credit to Robbie Williams and Guy Chambers. It\'s a song that will be p layed forever and bridges all age groups.\r\n\r\nRobbie did in no way deserve t hat mantle. Whenever we have these awards it is always \'artists\' from the pas t five years that seem to win the best of the best...We forget about the late 8 0s and early 90s for example. They weren\'t cool at the time, but because they are cool again now shows that the songs have greater longevity than people thin k.\r\n\r\nYes Angels is the best song since the past 25 years, because it touch es the soul as it carries a lot of meaning.\r\n\r\nI\'ve always disliked Angels intensely. I believe it to be symbolic of the general capacity of British pre-t eens, teens and middle aged women to accept low quality/ low aspirational music as "classic" songwriting. It\'s \'orrible. It seems obvious to me that people w ho like Robbie are people who don\'t particularly like music all that much. Fol ks without collections; folks who have never engaged in that madness one experi ences when falling under the spell of pop music. Angels adds nothing - it is me rely an irritating distraction - a wasp that refuses to go away on a summer\'s afternoon picnic. What a dreadful result. If you voted for it - you should feel ashamed of yourself - you probably only know a dozen songs or so don\'t you - s o where do get off applying this uninformed filter and casting this ridiculous vote. Booo hisss\r\n\r\nAngels, best song? You are kidding, right? Last five ye ars I might be willing to accept, but 25, no way. Did whoever voted for this ac tually have ANY music knowledge prior to, say, 1995? Really quite insulting to the British music industry of the past quarter of a century.\r\n\r\nNo surprise about Robbie Williams considering the list. Where on earth did the nominations list come from???? Compiled by an eleven year-old girl perhaps?? I mean, Will Y oung? Come on.\r\n\r\nWhat a load of crap, best song in the last 25 years - I d on\'t think so!! What about all the REAL artists out there over the last 25 yea rs - the list is endless, but Robbie Williams doesn\'t even come close.\r\n\r\n What a joke. That song has become such a bane to me that I have developed a Pav lovian response to the word \'Angels\' where I thrash around, and scream "no no no no no" until someone tells me "the radio\'s off". Why a half-baked cheesy di tty like Angels, which has become the anthem for millions of romantic sops (thi nk how many times it was sung on Pop Idol for example, and by whom), should be voted the best song of the past 25 years, is beyond me. If this is the song aga inst which all others are judged, then musicians may as well give up.\r\n\r\nWh y do we reward mediocrity so highly in this country?? The initial list was very weak anyway, but Angels the best song of the last 25 years!!!! I think not, I d idn\'t realize Radio 2 had so many listeners under the age of 10!!\r\n\r\nPeopl e have such short memories! A great song yes, but the best of the last 25 year s? Not a chance. I think the person as opposed to the song has been voted for h ere.\r\n\r\nThe Great British Public at work again. It\'s a mediocre, sentiment al and safe song. Granted, it\'s not too bad, either. But can it stand up again st ANYTHING by The Smiths (in particular "How Soon is Now?") or anything from t he Stone Roses\' first album? Nope.\r\n\r\nNo, Because I\'m not female and I\'m not 10!\r\n\r\nOk I like Robbie and Angels is a decent song. But it is no way t he best song of the past 25 years! The shortlist wasn\'t great but him winning it is a joke!\r\n\r\nPredictable and laughable.\r\n\r\nThe success of Angels at this years Brits reflects poorly on the state of British music over the last 25 years. The British public are brainwashed by the corporate pulp that is presnet ed to them as cutting edge music and true talent is being sadly missed. Whilst Angels is a popular song it is not even the best song in Robbie\'s repertoire n ever mind best song of the last 25 years.\r\n\r\nI am a huge Robbie fan and lov e that song. But I think there are a lot more outstanding songs / music out the re that influenced music today, unfortunately they were left out of the list.\r \n\r\nHave Radio 2 listeners even heard of Joy Division? A band who, through tw o albums, have had a bigger impact on music, and continue to do so, over the la st 25 years than Robbie Williams ever will.\r\n\r\nNo doubt about it. There\'s

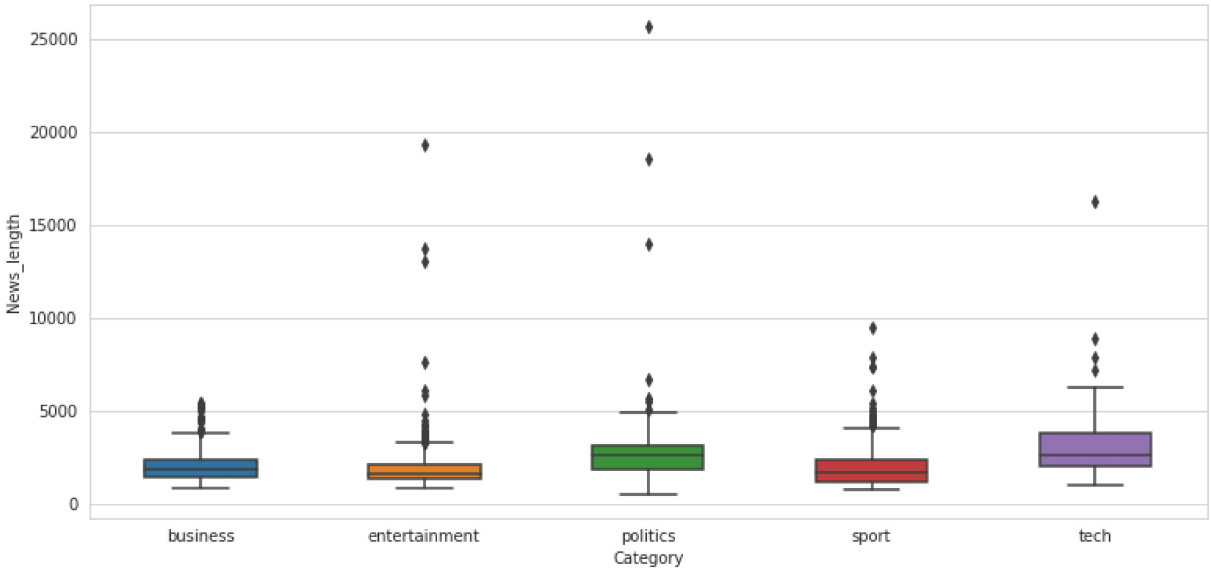
not a song done by anyone with more emotion and feeling. Some people will adopt their slightly snobby stances, but Angels has hit home with a far larger audien ce than any other song.\r\n\r\nIt should have been Joy Division.Those guys have played an influential part in shaping modern day music while Angels remains yet another pop song. I sincerely do not believe that in 25 years from now, the mos t influential artists will argue that Robbie Williams inspired their art in the way that the Byrds, the Beatles and Nick Drake have done for music today.\r\n\r \nYes!! I think its a brilliantly written song with different meanings to diffe rent people. There were other great songs in the category... but somebody had t o win! Robbie was a deserved winner.\r\n\r\nI am astounded that such a second-r ate record has beaten such a line up of amazing songs! it\'s a terrible song, v oted for by the masses who don\'t have the brains to appreciate innovative and exciting music.\r\n\r\nThe best song of the past 25 years? Really? Come on, thi s is surely a joke? No? I think I need a long lie down...\r\n\r\nIt just goes t o show that the british public do not have a clue about good quality music when they pick Robbie Williams over the beautifully talented Kate Bush and Joy Divis ion. I suppose it\'s confirmed one thing - the British public are consistently dull :- <(\r\n\r\nNo offence Robbie, but pleeease! There must be a thousand better songs than a formulated cheesy pop song for kids. No one agrees with thi s and quite frankly it is an embarassment to the integrity of British music, an d a further nail in the coffin!\r\n\r\nI think Robbie deserves it, he has been the most iconic of any stars we have had in Britain since John Lennon and is an idol to millions worldwide. Anybody who says he doesn\'t deserve it is jealous of his success. The only real challenger was Queen but hey, Another One Bites t he Dust!!\r\n\r\nObviously all the Karaoke singers in the UK voted for it.\r\n \r\nIf this is the best song of the last 25 years then the British Music indust ry 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 ty pe of music I normally like, but even as a diehard rock fan, I recognise that i t 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-rounder. Just like Robbie.\r\n\r\nB est song in 25 years? Since 1980? I\'m confused. "Angels" isn\'t a bad song. It \'s a nice, catchy, formulaic anthem that ticks all the boxes. But this is not great music. If anything it\'s regressive. Bland even. I suppose it\'s just mor e evidence of how redundant the Brit Awards have become.\r\n\r\nGranted angels is a good song, however it really wasn\'t up against any other proper competiti on. The Queen\'s song was lackluster, and apart from Kate Bush, the other choic es were pathetic! Also, why weren\'t the Stones there, David Bowie, etc, there are so many greater songs than Angels...I wonder if it was simply the fact that Robbie wasn\'t getting more awards so they had to make one up for him!\r\n\r\nB est song of the last 25 years? What a ridiculuous concept, and an even more rid iculous winner. Sigh. On the upside, at least it wasn\'t Bohemian Rhapsody, for which we should all be thankful.\r\n\r\nAngels is without doubt a great song bu t I really don\'t think it deserves the title of best song in 25 years.perhaps the vote had more to do with teenage opinion on Robbies\' goodlooks than the ac tual song!!! Don\'t get me wrong, I\'m not disputing his looks, but there are m ore deserved winners.\r\n\r\nIt would have been a travesty had Angels not won. Without Angels, Robbie Williams may well not be where he is now, and Britain wo uld have been deprived of one of its most charismatic and talented performing a rtists. It has to be seen performed live, with 125,000 people singing along to be fully appreciated. Well done Rob.\r\n\r\nI find it hard to believe that \'An gels\' is the best we have to show for the past 25 years! I\'m rapidly redifini ng \'best\' in my own head now to mean \'most gratuitously played at weddings a nd funerals because people think it has deep meaning\'. What about Britpop? Blu r, Oasis, Suede, Pulp... not only making fantastic songs but also making change s, doing something different. Why must \'best\' always come down to most commer cially popular?\r\n\r\nI\'ve nothing against Robbie, I actually like his music,

```
but how can this possibly be the best song from the last 25year? The Brits has
proved to be nothing more than a bargaining tool between the pop moguls to boos
t band profiles and record sales. The same goes with the Scissor Sisters, I thi
nk this is a superb record and thoroughly deserves the newcomer award, but the
album comes no where near U2\'s new record, neither are they in the same leagu
e. Once again there have been some baffling discisions made, they are not for a
rtisic reasons, but for profit.\r\n\r\nNo surprise really, it\'s voted for by t
he general public. Since when did they have taste in music?\r\n\r\nPersonally I
find Angels by Robbie Williams to be one of the most irritating songs I have ev
er heard!\r\n\r\nIt absolutely deserved to win. It is a song that has united th
e generations and will continue to be played for many years to come.\r\n\r\nIt
\'s an absolute joke, however most of the original 25 were very poor choices as
well. All in all a pretty pointless exercise !\r\n\r\nThe song is overplayed an
d oversentimental. Out of the rather poor five choices that were left, it shoul
d have gone to either Joy Division or Queen. I suppose we should be thankful th
at it didn\'t end up in the hands of Will Young though.\r\n\r\nAlthough it has
nostalgia value, there is no way it deserved to win. Everybody knows the words
to Bohemian Rhapsody, Nothing Compares 2 U, etc. Much better songs and more tim
eless. Give it to someone with real talent.\r\n\r\nAlthough Angels is a good so
ng I think that anyone with the slightest musical taste will realise that this
is not the best song of the last 25 years. This is just another example of reco
rd company manipulation to keep an artist in the public eye. Why not give him a
n award for the greatest pair of trousers if that\'s all it means!'
```

It's just a large news article.

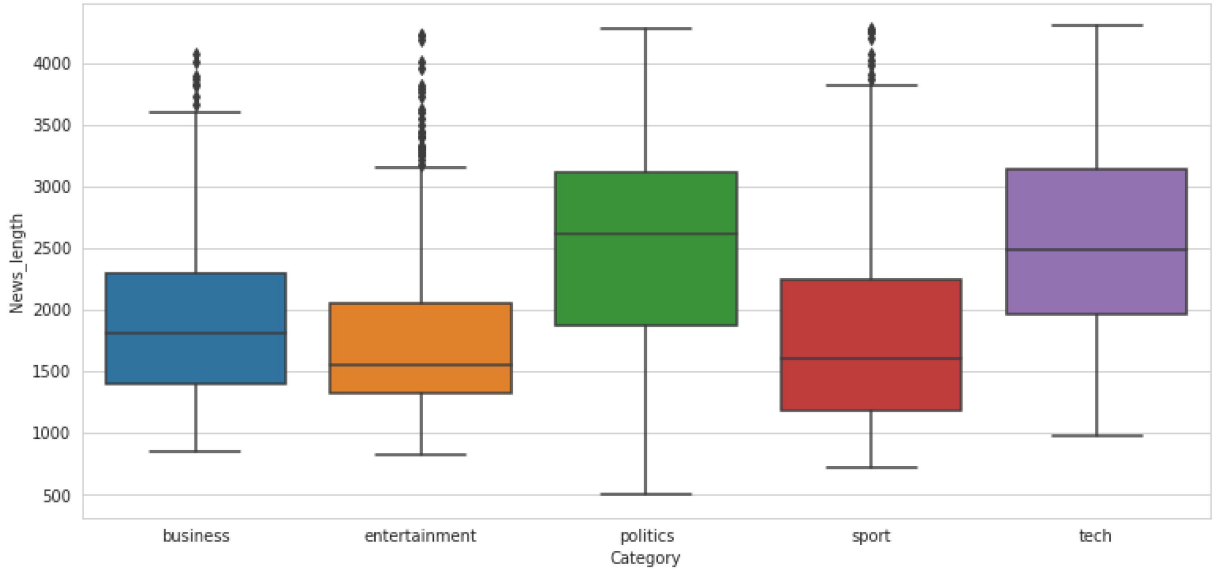
Let's now plot a boxplot:

In [13]:
```
plt.figure(figsize=(12.8,6))
sns.boxplot(data=df, x='Category', y='News_length', width=.5);
```



Now, let's remove the larger documents for better comprehension:

In [14]:
```
plt.figure(figsize=(12.8,6))
sns.boxplot(data=df_95, x='Category', y='News_length');
```



We can see that, although the length distribution is different for every category, the difference is not too big. If we had way too different lengths between categories we would have a problem since the feature creation process may take into account counts of words. However, when creating the features with TF-IDF scoring, we will normalize the features just to avoid this.

At this point, we cannot do further Exploratory Data Analysis. We'll turn onto the **Feature Engineering** section.

We'll save the dataset:

In [15]:
```
with open('News_dataset.pickle', 'wb') as output:
    pickle.dump(df, output)
```