

```
In [ ]: #This jupyter notebook is prepared by "Tulio Contramaestre"
```

## 1. Load data and perform general EDA

```
In [ ]:  
import numpy as np  
import matplotlib as mpl  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
import missingno as msno  
%matplotlib inline  
import scipy as stats  
import sklearn as sklrn  
from sklearn.model_selection import train_test_split  
from sklearn.preprocessing import StandardScaler  
from sklearn.linear_model import LinearRegression  
import sklearn.metrics as metrics  
from sklearn.linear_model import SGDRegressor  
from sklearn.preprocessing import PolynomialFeatures  
from sklearn.linear_model import Ridge  
from sklearn.linear_model import Lasso  
from sklearn.linear_model import ElasticNet  
from sklearn.preprocessing import MinMaxScaler  
from imblearn.over_sampling import SMOTE  
from sklearn.decomposition import PCA  
from sklearn.linear_model import LogisticRegression  
from sklearn.pipeline import Pipeline  
from sklearn.model_selection import RepeatedStratifiedKFold  
from sklearn.model_selection import cross_val_score  
from sklearn.metrics import confusion_matrix  
from sklearn.model_selection import cross_val_predict  
from sklearn.svm import SVC  
from sklearn.metrics import plot_confusion_matrix  
from sklearn.neighbors import KNeighborsClassifier  
from sklearn.metrics import classification_report  
from sklearn.naive_bayes import GaussianNB  
from sklearn.model_selection import RepeatedStratifiedKFold, GridSearchCV  
from sklearn.naive_bayes import CategoricalNB  
from sklearn.svm import SVC  
from sklearn import tree  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.ensemble import GradientBoostingClassifier  
from sklearn.ensemble import GradientBoostingRegressor  
from sklearn.ensemble import AdaBoostClassifier  
from sklearn.metrics import PrecisionRecallDisplay  
from sklearn.metrics import precision_recall_curve  
import scipy.cluster.hierarchy as shc  
from sklearn.cluster import AgglomerativeClustering  
from sklearn.preprocessing import normalize  
import nltk  
from wordcloud import WordCloud  
from sklearn.feature_extraction.text import CountVectorizer  
from sklearn.feature_extraction.text import TfidfTransformer  
from nltk.corpus import stopwords  
from nltk.stem.lancaster import LancasterStemmer  
import string
```

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.neural_network import MLPClassifier
import tensorflow as tf
```

In [ ]:

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

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

Out[ ]:

In [ ]:

```
text = [line.rstrip() for line in open("news.csv")]
print(len(text))
```

44955

In [ ]:

```
for message in text[:10]:
    print(message)
    print('\n')
```

title text subject date target

As U.S. budget fight looms, Republicans flip their fiscal script WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts, called himself a "fiscal conservative" on Sunday and urged budget restraint in 2018. In keeping with a sharp pivot under way among Republicans, U.S. Representative Mark Meadows, speaking on CBS' "Face the Nation," drew a hard line on federal spending, which lawmakers are bracing to do battle over in January. When they return from the holidays on Wednesday, lawmakers will begin trying to pass a federal budget in a fight likely to be linked to other issues, such as immigration policy, even as the November congressional election campaigns approach in which Republicans will seek to keep control of Congress. President Donald Trump and his Republicans want a big budget increase in military spending, while Democrats also want proportional increases for non-defense "discretionary" spending on programs that support education, scientific research, infrastructure, public health and environmental protection. "The (Trump) administration has already been willing to say: 'We're going to increase non-defense discretionary spending ... by about 7 percent,'" Meadows, chairman of the small but influential House Freedom Caucus, said on the program. "Now, Democrats are saying that's not enough, we need to give the government a pay raise of 10 to 11 percent. For a fiscal conservative, I don't see where the rationale is. ... Eventually you run out of other people's money," he said. Meadows was among Republicans who voted in late December for their party's debt-financed tax overhaul, which is expected to balloon t

he federal budget deficit and add about \$1.5 trillion over 10 years to the \$20 trillion national debt. "It's interesting to hear Mark talk about fiscal responsibility," Democratic U.S. Representative Joseph Crowley said on CBS. Crowley said the Republican tax bill would require the United States to borrow \$1.5 trillion, to be paid off by future generations, to finance tax cuts for corporations and the rich. "This is one of the least ... fiscally responsible bills we've ever seen passed in the history of the House of Representatives. I think we're going to be paying for this for many, many years to come," Crowley said. Republicans insist the tax package, the biggest U.S. tax overhaul in more than 30 years, will boost the economy and job growth. House Speaker Paul Ryan, who also supported the tax bill, recently went further than Meadows, making clear in a radio interview that welfare or "entitlement reform," as the party often calls it, would be a top Republican priority in 2018. In Republican parlance, "entitlement" programs mean food stamps, housing assistance, Medicare and Medicaid health insurance for the elderly, poor and disabled, as well as other programs created by Washington to assist the needy. Democrats seized on Ryan's early December remarks, saying they showed Republicans would try to pay for their tax overhaul by seeking spending cuts for social programs. But the goals of House Republicans may have to take a back seat to the Senate, where the votes of some Democrats will be needed to approve a budget and prevent a government shutdown. Democrats will use their leverage in the Senate, which Republicans narrowly control, to defend both discretionary non-defense programs and social spending, while tackling the issue of the "Dreamers," people brought illegally to the country as children. Trump in September put a March 2018 expiration date on the Deferred Action for Childhood Arrivals, or DACA, program, which protects the young immigrants from deportation and provides them with work permits. The president has said in recent Twitter messages he wants funding for his proposed Mexican border wall and other immigration law changes in exchange for agreeing to help the Dreamers. Representative Debbie Dingell told CBS she did not favor linking that issue to other policy objectives, such as wall funding. "We need to do DACA clean," she said. On Wednesday, Trump aides will meet with congressional leaders to discuss those issues. That will be followed by a weekend of strategy sessions for Trump and Republican leaders on Jan. 6 and 7, the White House said. Trump was also scheduled to meet on Sunday with Florida Republican Governor Rick Scott, who wants more emergency aid. The House has passed an \$81 billion aid package after hurricanes in Florida, Texas and Puerto Rico, and wildfires in California. The package far exceeded the \$44 billion requested by the Trump administration. The Senate has not yet voted on the aid.

politicsNews December 31, 2017

1

U.S. military to accept transgender recruits on Monday: Pentagon WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. military starting on Monday as ordered by federal courts, the Pentagon said on Friday, after President Donald Trump's administration decided not to appeal rulings that blocked his transgender ban. Two federal appeals courts, one in Washington and one in Virginia, last week rejected the administration's request to put on hold orders by lower court judges requiring the military to begin accepting transgender recruits on Jan. 1. A Justice Department official said the administration will not challenge those rulings. "The Department of Defense has announced that it will be releasing an independent study of these issues in the coming weeks. So rather than litigate this interim appeal before that occurs, the administration has decided to wait for DOD's study and will continue to defend the president's lawful authority in District Court in the meantime," the official said, speaking on condition of anonymity. In September, the Pentagon said it had created a panel of senior officials to study how to implement a directive by Trump to prohibit transgender individuals from serving. The Defense Department has until Feb. 21 to submit a plan to Trump. Lawyers representing currently-serving transgender service members and aspiring recruits said they had expected the administration to appeal the rulings to the conservative-majority Supreme Court, but were hoping that would not happen. Pentagon spokeswoman Heather Babb said in a statement: "As mandated by court order, the Department of Defense is prepared to begin accessing transgender applicants for military service Jan. 1. All applicants must meet all accession standards." Jennifer Levi, a lawyer with gay, lesbian and transgender advocacy group GLAD, called the decision not to appeal "great news." "I'm hoping it means the government has come to see that there is no way to justify a ban and t

hat it's not good for the military or our country," Levi said. Both GLAD and the American Civil Liberties Union represent plaintiffs in the lawsuits filed against the administration. In a move that appealed to his hard-line conservative supporters, Trump announced in July that he would prohibit transgender people from serving in the military, reversing Democratic President Barack Obama's policy of accepting them. Trump said on Twitter at the time that the military "cannot be burdened with the tremendous medical costs and disruption that transgender in the military would entail." Four federal judges - in Baltimore, Washington, D.C., Seattle and Riverside, California - have issued rulings blocking Trump's ban while legal challenges to the Republican president's policy proceed. The judges said the ban would likely violate the right under the U.S. Constitution to equal protection under the law. The Pentagon on Dec. 8 issued guidelines to recruitment personnel in order to enlist transgender applicants by Jan. 1. The memo outlined medical requirements and specified how the applicants' sex would be identified and even which uniforms they would wear. The Trump administration previously said in legal papers that the armed forces were not prepared to train thousands of personnel on the medical standards needed to process transgender applicants and might have to accept "some individuals who are not medically fit for service." The Obama administration had set a deadline of July 1, 2017, to begin accepting transgender recruits. But Trump's defense secretary, James Mattis, postponed that date to Jan. 1, 2018, which the president's ban then put off indefinitely. Trump has taken other steps aimed at rolling back transgender rights. In October, his administration said a federal law banning gender-based workplace discrimination does not protect transgender employees, reversing another Obama-era position. In February, Trump rescinded guidance issued by the Obama administration saying that public schools should allow transgender students to use the restroom that corresponds to their gender identity.

politicsNews December 29, 2017

1

Senior U.S. Republican senator: 'Let Mr. Mueller do his job' WASHINGTON (Reuters) - The special counsel investigation of links between Russia and President Trump's 2016 election campaign should continue without interference in 2018, despite calls from some Trump administration allies and Republican lawmakers to shut it down, a prominent Republican senator said on Sunday. Lindsey Graham, who serves on the Senate armed forces and judiciary committees, said Department of Justice Special Counsel Robert Mueller needs to carry on with his Russia investigation without political interference. "This investigation will go forward. It will be an investigation conducted without political influence," Graham said on CBS's Face the Nation news program. "And we all need to let Mr. Mueller do his job. I think he's the right guy at the right time." The question of how Russia may have interfered in the election, and how Trump's campaign may have had links with or co-ordinated any such effort, has loomed over the White House since Trump took office in January. It shows no sign of receding as Trump prepares for his second year in power, despite intensified rhetoric from some Trump allies in recent weeks accusing Mueller's team of bias against the Republican president. Trump himself seemed to undercut his supporters in an interview last week with the New York Times in which he said he expected Mueller was "going to be fair." Russia's role in the election and the question of possible links to the Trump campaign are the focus of multiple inquiries in Washington. Three committees of the Senate and the House of Representatives are investigating, as well as Mueller, whose team in May took over an earlier probe launched by the U.S. Federal Bureau of Investigation (FBI). Several members of the Trump campaign and administration have been convicted or indicted in the investigation. Trump and his allies deny any collusion with Russia during the campaign, and the Kremlin has denied meddling in the election. Graham said he still wants an examination of the FBI's use of a dossier on links between Trump and Russia that was compiled by a former British spy, Christopher Steele, which prompted Trump allies and some Republicans to question Mueller's inquiry. On Saturday, the New York Times reported that it was not that dossier that triggered an early FBI probe, but a tip from former Trump campaign foreign policy adviser George Papadopoulos to an Australian diplomat that Russia had damaging information about former Trump rival Hillary Clinton. "I want somebody to look at the way the Department of Justice used this dossier. It bothers me greatly the way they used it, and I want somebody to look at it," Graham said. But he said the Russia investigation must continue. "As a matter of fact, it would hurt us if we ignored it," he said.

politicsNews December 31, 2017

1

FBI Russia probe helped by Australian diplomat tip-off: NYT      WASHINGTON (Reuters) - Trump campaign adviser George Papadopoulos told an Australian diplomat in May 2016 that Russia had political dirt on Democratic presidential candidate Hillary Clinton, the New York Times reported on Saturday. The conversation between Papadopoulos and the diplomat, Alexander Downer, in London was a driving factor behind the FBI's decision to open a counter-intelligence investigation of Moscow's contacts with the Trump campaign, the Times reported. Two months after the meeting, Australian officials passed the information that came from Papadopoulos to their American counterparts when leaked Democratic emails began appearing online, according to the newspaper, which cited four current and former U.S. and foreign officials. Besides the information from the Australians, the probe by the Federal Bureau of Investigation was also propelled by intelligence from other friendly governments, including the British and Dutch, the Times said. Papadopoulos, a Chicago-based international energy lawyer, pleaded guilty on Oct. 30 to lying to FBI agents about contacts with people who claimed to have ties to top Russian officials. It was the first criminal charge alleging links between the Trump campaign and Russia. The White House has played down the former aide's campaign role, saying it was "extremely limited" and that any actions he took would have been on his own. The New York Times, however, reported that Papadopoulos helped set up a meeting between then-candidate Donald Trump and Egyptian President Abdel Fattah al-Sisi and edited the outline of Trump's first major foreign policy speech in April 2016. The federal investigation, which is now being led by Special Counsel Robert Mueller, has hung over Trump's White House since he took office almost a year ago. Some Trump allies have recently accused Mueller's team of being biased against the Republican president. Lawyers for Papadopoulos did not immediately respond to requests by Reuters for comment. Mueller's office declined to comment. Trump's White House attorney, Ty Cobb, declined to comment on the New York Times report. "Out of respect for the special counsel and his process, we are not commenting on matters such as this," he said in a statement. Mueller has charged four Trump associates, including Papadopoulos, in his investigation. Russia has denied interfering in the U.S. election and Trump has said there was no collusion between his campaign and Moscow.

politicsNews      December

30, 2017

1

Trump wants Postal Service to charge 'much more' for Amazon shipments      SEATTLE/WASHINGTON (Reuters) - President Donald Trump called on the U.S. Postal Service on Friday to charge "much more" to ship packages for Amazon (AMZN.O), picking another fight with an online retail giant he has criticized in the past. "Why is the United States Post Office, which is losing many billions of dollars a year, while charging Amazon and others so little to deliver their packages, making Amazon richer and the Post Office dumber and poorer? Should be charging MUCH MORE!" Trump wrote on Twitter. The president's tweet drew fresh attention to the fragile finances of the Postal Service at a time when tens of millions of parcels have just been shipped all over the country for the holiday season. The U.S. Postal Service, which runs at a big loss, is an independent agency within the federal government and does not receive tax dollars for operating expenses, according to its website. Package delivery has become an increasingly important part of its business as the Internet has led to a sharp decline in the amount of first-class letters. The president does not determine postal rates. They are set by the Postal Regulatory Commission, an independent government agency with commissioners selected by the president from both political parties. That panel raised prices on packages by almost 2 percent in November. Amazon was founded by Jeff Bezos, who remains the chief executive officer of the retail company and is the richest person in the world, according to Bloomberg News. Bezos also owns The Washington Post, a newspaper Trump has repeatedly railed against in his criticisms of the news media. In tweets over the past year, Trump has said the "Amazon Washington Post" fabricated stories. He has said Amazon does not pay sales tax, which is not true, and so hurts other retailers, part of a pattern by the former businessman and reality television host of periodically turning his ire on big American companies since he took office in January. Daniel Ives, a research analyst at GBH Insights, said Trump's comment could be taken as a warning to the retail giant. However, he said he was not concerned for Amazon. "We do not see any price hikes in the future. However, that is a risk that Am

azon is clearly aware of and (it) is building out its distribution (system) aggressively," he said. Amazon has shown interest in the past in shifting into its own delivery service, including testing drones for deliveries. In 2015, the company spent \$11.5 billion on shipping, 46 percent of its total operating expenses that year. Amazon shares were down 0.86 percent to \$1,175.90 by early afternoon. Overall, U.S. stock prices were down slightly on Friday. Satish Jindel, president of ShipMatrix Inc, which analyzes shipping data, disputed the idea that the Postal Service charges less than United Parcel Service Inc (UPS.N) and FedEx Corp (FDX.N), the other biggest players in the parcel delivery business in the United States. Many customers get lower rates from UPS and FedEx than they would get from the post office for comparable services, he said. The Postal Service delivers about 62 percent of Amazon packages, for about 3.5 to 4 million a day during the current peak year-end holiday shipping season, Jindel said. The Seattle-based company and the post office have an agreement in which mail carriers take Amazon packages on the last leg of their journeys, from post offices to customers' doorsteps. Amazon's No. 2 carrier is UPS, at 21 percent, and FedEx is third, with 8 percent or so, according to Jindel. Trump's comment tapped into a debate over whether Postal Service pricing has kept pace with the rise of e-commerce, which has flooded the mail with small packages. Private companies like UPS have long claimed the current system unfairly undercuts their business. Steve Gaut, a spokesman for UPS, noted that the company values its "productive relationship" with the postal service, but that it has filed with the Postal Regulatory Commission its concerns about the postal service's methods for covering costs. Representatives for Amazon, the White House, the U.S. Postal Service and FedEx declined comment or were not immediately available for comment on Trump's tweet. According to its annual report, the Postal Service lost \$2.74 billion this year, and its deficit has ballooned to \$61.86 billion. While the Postal Service's revenue for first class mail, marketing mail and periodicals is flat or declining, revenue from package delivery is up 44 percent since 2014 to \$19.5 billion in the fiscal year ended Sept. 30, 2017. But it also lost about \$2 billion in revenue when a temporary surcharge expired in April 2016. According to a Government Accountability Office report in February, the service is facing growing personnel expenses, particularly \$73.4 billion in unfunded pension and benefits liabilities. The Postal Service has not announced any plans to cut costs. By law, the Postal Service has to set prices for package delivery to cover the costs attributable to that service. But the postal service allocates only 5.5 percent of its total costs to its business of shipping packages even though that line of business is 28 percent of its total revenue.

News December 29, 2017 1

White House, Congress prepare for talks on spending, immigration WEST PALM BEACH, Fla./WASHINGTON (Reuters) - The White House said on Friday it was set to kick off talks next week with Republican and Democratic congressional leaders on immigration policy, government spending and other issues that need to be wrapped up early in the new year. The expected flurry of legislative activity comes as Republicans and Democrats begin to set the stage for midterm congressional elections in November. President Donald Trump's Republican Party is eager to maintain control of Congress while Democrats look for openings to wrest seats away in the Senate and the House of Representatives. On Wednesday, Trump's budget chief Mick Mulvaney and legislative affairs director Marc Short will meet with Senate Majority Leader Mitch McConnell and House Speaker Paul Ryan - both Republicans - and their Democratic counterparts, Senator Chuck Schumer and Representative Nancy Pelosi, the White House said. That will be followed up with a weekend of strategy sessions for Trump, McConnell and Ryan on Jan. 6 and 7 at the Camp David presidential retreat in Maryland, according to the White House. The Senate returns to work on Jan. 3 and the House on Jan. 8. Congress passed a short-term government funding bill last week before taking its Christmas break, but needs to come to an agreement on defense spending and various domestic programs by Jan. 19, or the government will shut down. Also on the agenda for lawmakers is disaster aid for people hit by hurricanes in Puerto Rico, Texas and Florida, and by wildfires in California. The House passed an \$81 billion package in December, which the Senate did not take up. The White House has asked for a smaller figure, \$44 billion. Deadlines also loom for soon-to-expire protections for young adult immigrants who entered the country illegally as children, known as "Dreamers." In September, Trump ended Democratic former President Barack Obama's Deferred Action for Childhood Arrivals (DACA)

program, which protected Dreamers from deportation and provided work permits, effective in March, giving Congress until then to devise a long-term solution. Democrats, some Republicans and a number of large companies have pushed for DACA protections to continue. Trump and other Republicans have said that will not happen without Congress approving broader immigration policy changes and tougher border security. Democrats oppose funding for a wall promised by Trump along the U.S.-Mexican border. "The Democrats have been told, and fully understand, that there can be no DACA without the desperately needed WALL at the Southern Border and an END to the horrible Chain Migration & ridiculous Lottery System of Immigration etc," Trump said in a Twitter post on Friday. Trump wants to overhaul immigration rules for extended families and others seeking to live in the United States. Republican U.S. Senator Jeff Flake, a frequent critic of the president, said he would work with Trump to protect Dreamers. "We can fix DACA in a way that beefs up border security, stops chain migration for the DREAMers, and addresses the unfairness of the diversity lottery. If POTUS (Trump) wants to protect these kids, we want to help him keep that promise," Flake wrote on Twitter. Congress in early 2018 also must raise the U.S. debt ceiling to avoid a government default. The U.S. Treasury would exhaust all of its borrowing options and run dry of cash to pay its bills by late March or early April if Congress does not raise the debt ceiling before then, according to the nonpartisan Congressional Budget Office. Trump, who won his first major legislative victory with the passage of a major tax overhaul this month, has also promised a major infrastructure plan. politics News December 29, 2017 1

Trump says Russia probe will be fair, but timeline unclear: NYT WEST PALM BEACH, Fla (Reuters) - President Donald Trump said on Thursday he believes he will be fairly treated in a special counsel investigation into Russian meddling in the U.S. presidential election, but said he did not know how long the probe would last. The federal investigation has hung over Trump's White House since he took office almost a year ago, and some Trump allies have in recent weeks accused the team of Justice Department Special Counsel Robert Mueller of being biased against the Republican president. But in an interview with the New York Times, Trump appeared to shrug off concerns about the investigation, which was prompted by U.S. intelligence agencies' conclusion that Russia tried to help Trump defeat Democrat Hillary Clinton by hacking and releasing embarrassing emails and disseminating propaganda. "There's been no collusion. But I think he's going to be fair," Trump said in what the Times described as a 30-minute impromptu interview at his golf club in West Palm Beach, Florida. Mueller has charged four Trump associates in his investigation. Russia has denied interfering in the U.S. election. U.S. Deputy Attorney General Rod Rosenstein said this month that he was not aware of any impropriety by Mueller's team. Trump's lawyers have been saying for weeks that they had expected the Mueller investigation to wrap up quickly, possibly by the end of 2017. Mueller has not commented on how long it will last. Trump told the Times that he did not know how long the investigation would take. "Timing-wise, I can't tell you. I just don't know," he said. Trump said he thought a prolonged probe "makes the country look bad" but said it has energized his core supporters. "What it's done is, it's really angered the base and made the base stronger. My base is strong than it's ever been," he said. The interview was a rare break in Trump's Christmas vacation in Florida. He has golfed each day aside from Christmas Day, and mainly kept a low profile, apart from the occasional flurry of tweets. He spent one day golfing with Republican Senator David Perdue from Georgia, who has pushed legislation to cap immigration numbers, and had dinner on Thursday with Commerce Secretary Wilbur Ross, an international trade hawk. Trump told the Times he hoped to work with Democrats in the U.S. Congress on a spending plan to fix roads and other infrastructure, and on protections for a group of undocumented immigrants who were brought to the United States as children. Trump spoke about trade issues, saying he had backed off his hard line on Chinese trade practices in the hope that Beijing would do more to pressure North Korea to end its nuclear and missile testing program. He said he had been disappointed in the results. He also complained about the North American Free Trade Agreement (NAFTA), which his administration is attempting to renegotiate in talks with Mexico and Canada. Trump said Canadian Prime Minister Justin Trudeau had played down the importance of Canadian oil and lumber exports to the United States when looking at the balance of trade between the two countries. "If I don't make the right deal, I'll terminate NAFTA in two seconds. But we're doing

pretty good," Trump said.

politicsNews December 29, 2017

1

Factbox: Trump on Twitter (Dec 29) - Approval rating, Amazon  
 The following statements were posted to the verified Twitter accounts of U.S. President Donald Trump, @realDonaldTrump and @POTUS. The opinions expressed are his own. Reuters has not edited the statements or confirmed their accuracy. @realDonaldTrump : - While the Fake News loves to talk about my so-called low approval rating, @foxandfriends just showed that my rating on Dec. 28, 2017, was approximately the same as President Obama on Dec. 28, 2009, which was 47%...and this despite massive negative Trump coverage & Russia hoax! [0746 EST] - Why is the United States Post Office, which is losing many billions of dollars a year, while charging Amazon and others so little to deliver their packages, making Amazon richer and the Post Office dumber and poorer? Should be charging MUCH MORE! [0804 EST] -- Source link: (bit.ly/2jBh4LU) (bit.ly/2jpEXYR) politicsNews December 29, 2017 1

Trump on Twitter (Dec 28) - Global Warming  
 The following statements were posted to the verified Twitter accounts of U.S. President Donald Trump, @realDonaldTrump and @POTUS. The opinions expressed are his own. Reuters has not edited the statements or confirmed their accuracy. @realDonaldTrump : - Together, we are MAKING AMERICA GREAT AGAIN! bit.ly/2lnpKaq [1814 EST] - In the East, it could be the COLDEST New Year's Eve on record. Perhaps we could use a little bit of that good old Global Warming that our Country, but not other countries, was going to pay TRILLIONS OF DOLLARS to protect against. Bundle up! [1901 EST] -- Source link: (bit.ly/2jBh4LU) (bit.ly/2jpEXYR) politicsNews December 29, 2017 1

```
In [ ]: data = pd.read_csv("news.csv", sep = '\t')
```

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

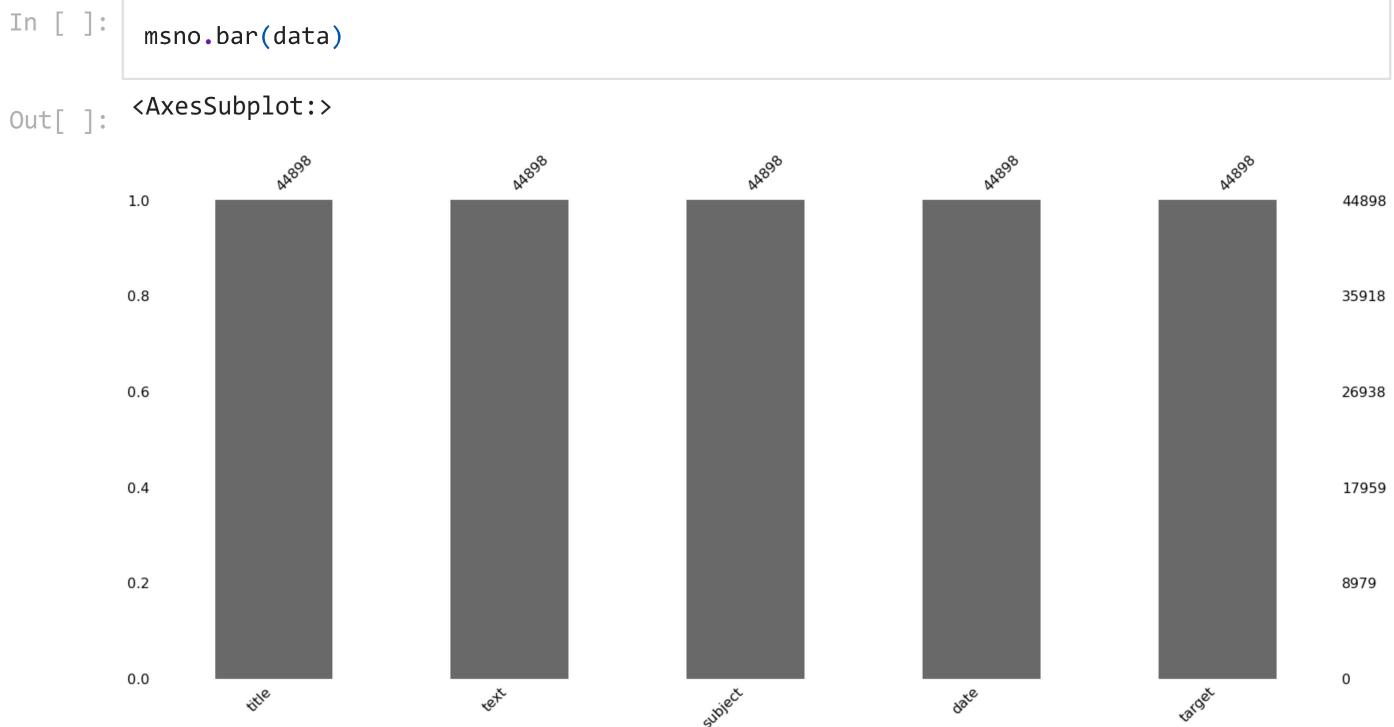
	title	text	subject	date	target
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

```
In [ ]: data.tail()
```

	title	text	subject	date	target
44893	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0

	<b>title</b>	<b>text</b>	<b>subject</b>	<b>date</b>	<b>target</b>
<b>44894</b>	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It's a familiar theme. ...	Middle-east	January 16, 2016	0
<b>44895</b>	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
<b>44896</b>	How to Blow \$700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
<b>44897</b>	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

## IV. Null Check



In [ ]:

```
data.isnull().sum()
```

Out[ ]:

```
title      0
text       0
subject    0
date       0
target     0
dtype: int64
```

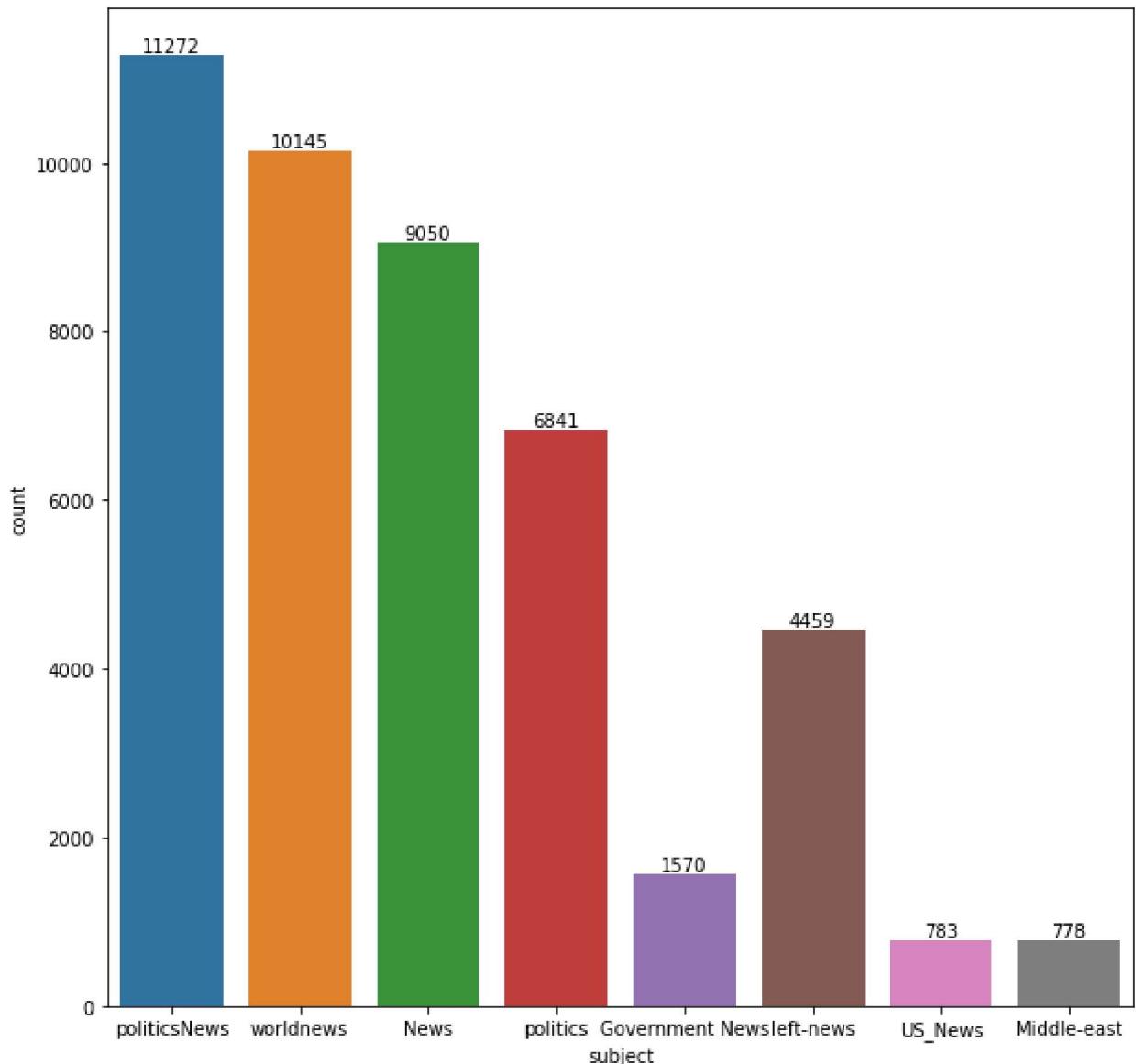
No null values

## V. Counterplot

In [ ]:

```
plt.figure(figsize= (10, 10))
plot = sns.countplot(x = 'subject', data= data)
plot.bar_label(plot.containers[0])
plt.show
```

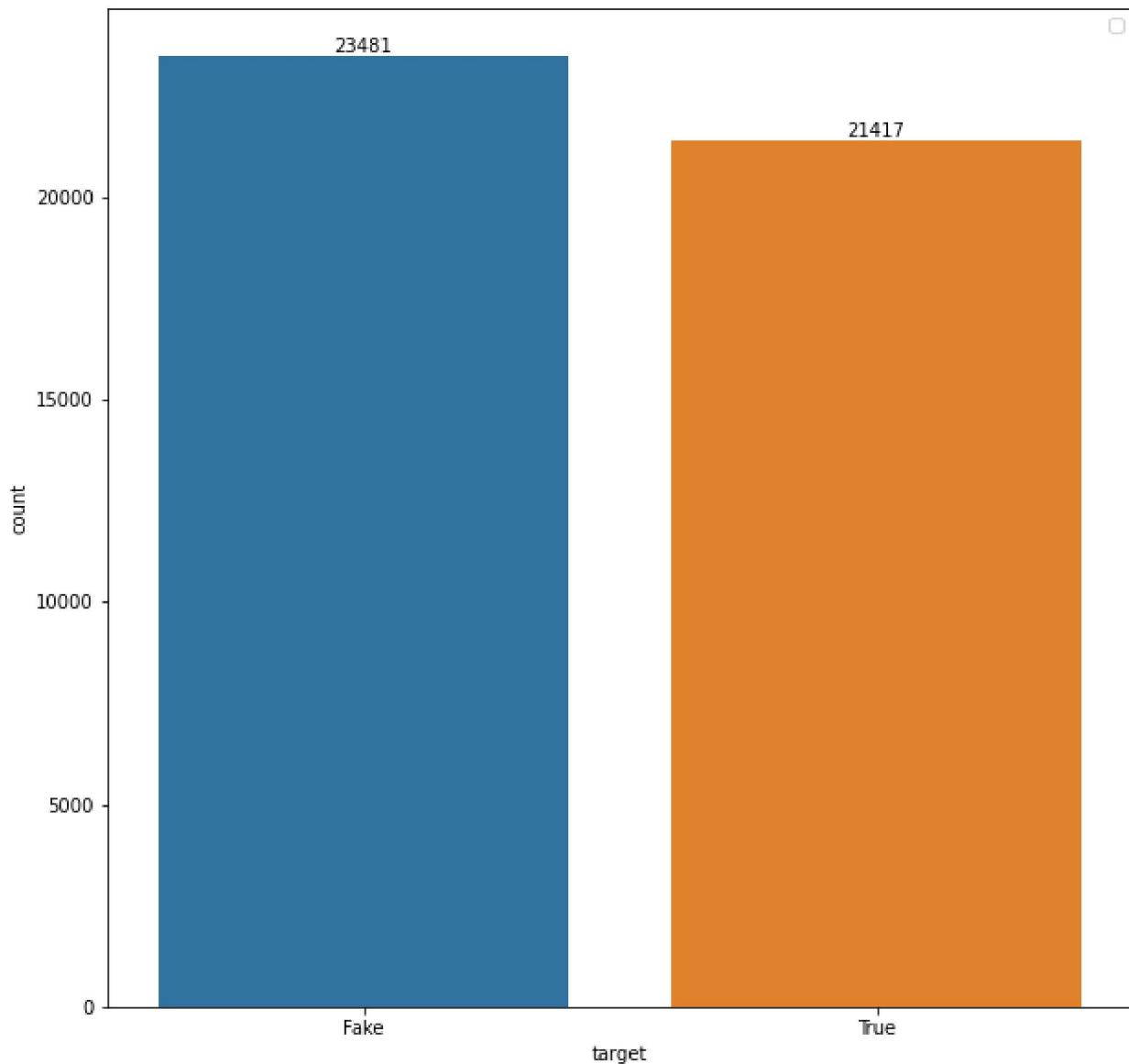
```
Out[ ]: <function matplotlib.pyplot.show(close=None, block=None)>
```



```
In [ ]: plt.figure(figsize= (10, 10))
plot = sns.countplot(x = 'target', data= data)
plot.bar_label(plot.containers[0])
plot.legend(loc = "upper right")
plot.set_xticklabels(["Fake", "True"])
plt.show
```

No handles with labels found to put in legend.

```
Out[ ]: <function matplotlib.pyplot.show(close=None, block=None)>
```



## VII. word clouds

### Fake news

```
In [ ]: falseNews = data.loc[data['target'] == 0, ['text']]
```

```
In [ ]: falseNews
```

```
Out[ ]:          text
```

**21417** Donald Trump just couldn't wish all Americans ...

**21418** House Intelligence Committee Chairman Devin Nu...

**21419** On Friday, it was revealed that former Milwauk...

**21420** On Christmas day, Donald Trump announced that ...

**21421** Pope Francis used his annual Christmas Day mes...

...

...

text

- 44893** 21st Century Wire says As 21WIRE reported earl...

**44894** 21st Century Wire says It s a familiar theme. ....

**44895** Patrick Henningsen 21st Century WireRemember ...

**44896** 21st Century Wire says Al Jazeera America will...

**44897** 21st Century Wire says As 21WIRE predicted in ...

23481 rows × 1 columns

In [ ]:

```
fakeNewsText = " ".join(text for text in falseNews['text'])
```

In [1]:

```
word_cloud = WordCloud(collocations= False, background_color= 'white').generate(fakeNews)
plt.figure(figsize= (10, 10))
plt.imshow(word_cloud, interpolation= 'bilinear')
plt.axis("off")
plt.show()
```



As we can see above the fake news word cloud focuses highly on the president and another surrounding names such as clinton or obama. New and will also seem to be decently prevalent. This could be due to the fact that these words can attract more readers.

## True news

In [ ]:

```
trueNews = data.loc[data['target'] == 1, ['text']]
```

In [ ]:

trueNews

Out[ ]:

text

- 0** WASHINGTON (Reuters) - The head of a conservat...

text

- 1** WASHINGTON (Reuters) - Transgender people will...
  - 2** WASHINGTON (Reuters) - The special counsel inv...
  - 3** WASHINGTON (Reuters) - Trump campaign adviser ...
  - 4** SEATTLE/WASHINGTON (Reuters) - President Donald...
  - ...
  - 21412** BRUSSELS (Reuters) - NATO allies on Tuesday we...
  - 21413** LONDON (Reuters) - LexisNexis, a provider of I...
  - 21414** MINSK (Reuters) - In the shadow of disused Sov...
  - 21415** MOSCOW (Reuters) - Vatican Secretary of State ...
  - 21416** JAKARTA (Reuters) - Indonesia will buy 11 Sukhoi...

21417 rows × 1 columns

```
In [ ]: trueNewsText = " ".join(text for text in trueNews['text'])
```

```
In [ ]: word_cloud = WordCloud(collocations= False, background_color= 'white').generate(trueNew  
plt.figure(figsize= (10, 10))  
plt.imshow(word_cloud, interpolation= 'bilinear')  
plt.axis("off")  
plt.show()
```



As we can see for the True news words such as "Said", "state", "government" are a bit more prominent in comparison to fake news text. This could be related to the idea of quoting people or researchers, since the articles tend to be a bit more credible when it is considered "true" news.

## VIII. "Alltext" column

In [ ]:

```
cols = ['title', 'text', 'subject']
data['AllText'] = data[cols].apply(lambda row: ' '.join(row.values.astype(str)), axis=1)
```

```
In [ ]: # data['AllText'] = data['title'] + " " + data['text'] + " " + data['subject']
```

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

	title	text	subject	date	target	AllText
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1	As U.S. budget fight looms, Republicans flip t...
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1	U.S. military to accept transgender recruits o...
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1	Senior U.S. Republican senator: 'Let Mr. Muell...
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1	FBI Russia probe helped by Australian diplomat...
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donald...	politicsNews	December 29, 2017	1	Trump wants Postal Service to charge 'much mor...

## IX. Copy

```
In [ ]: data_copy = data.copy()
```

## X. drop title,text,subject

```
In [ ]: data = data.drop(['text', 'title', 'subject', 'date'], axis= 1)
```

```
In [ ]: data
```

	target	AllText
0	1	As U.S. budget fight looms, Republicans flip t...
1	1	U.S. military to accept transgender recruits o...
2	1	Senior U.S. Republican senator: 'Let Mr. Muell...
3	1	FBI Russia probe helped by Australian diplomat...

	<b>target</b>	<b>AllText</b>
<b>4</b>	1	Trump wants Postal Service to charge 'much mor...
...	...	...
<b>44893</b>	0	McPain: John McCain Furious That Iran Treated ...
<b>44894</b>	0	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...
<b>44895</b>	0	Sunnistan: US and Allied 'Safe Zone' Plan to T...
<b>44896</b>	0	How to Blow \$700 Million: Al Jazeera America F...
<b>44897</b>	0	10 U.S. Navy Sailors Held by Iranian Military ...

44898 rows × 2 columns

## Xi. calculate the length

In [ ]: `data['AllText']`

Out[ ]: 0 As U.S. budget fight looms, Republicans flip t...  
 1 U.S. military to accept transgender recruits o...  
 2 Senior U.S. Republican senator: 'Let Mr. Muell...  
 3 FBI Russia probe helped by Australian diplomat...  
 4 Trump wants Postal Service to charge 'much mor...  
 ...  
 44893 McPain: John McCain Furious That Iran Treated ...  
 44894 JUSTICE? Yahoo Settles E-mail Privacy Class-ac...  
 44895 Sunnistan: US and Allied 'Safe Zone' Plan to T...  
 44896 How to Blow \$700 Million: Al Jazeera America F...  
 44897 10 U.S. Navy Sailors Held by Iranian Military ...  
 Name: AllText, Length: 44898, dtype: object

In [ ]: `data['length'] = data['AllText'].apply(len)`

In [ ]: `data`

	<b>target</b>	<b>AllText</b>	<b>length</b>
<b>0</b>	1	As U.S. budget fight looms, Republicans flip t...	4737
<b>1</b>	1	U.S. military to accept transgender recruits o...	4155
<b>2</b>	1	Senior U.S. Republican senator: 'Let Mr. Muell...	2863
<b>3</b>	1	FBI Russia probe helped by Australian diplomat...	2534
<b>4</b>	1	Trump wants Postal Service to charge 'much mor...	5287
...	...	...	...
<b>44893</b>	0	McPain: John McCain Furious That Iran Treated ...	3311
<b>44894</b>	0	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	1778
<b>44895</b>	0	Sunnistan: US and Allied 'Safe Zone' Plan to T...	25163

target	AllText	length
44896	0	How to Blow \$700 Million: Al Jazeera America F...
44897	0	10 U.S. Navy Sailors Held by Iranian Military ...

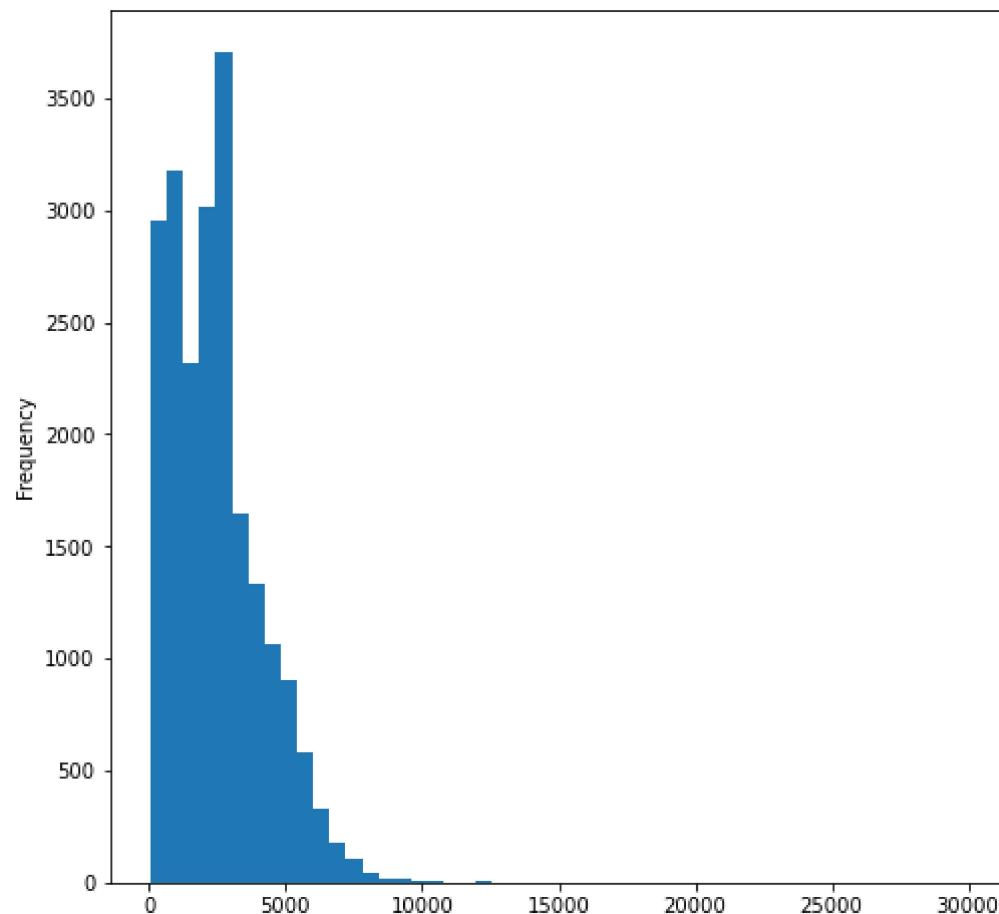
44898 rows × 3 columns

## XII. histograms for length

### True news

```
In [ ]: plt.figure(figsize= (8, 8))
trueNewsLength = data.loc[data['target'] == 1, ['length']]
trueNewsLength['length'].plot(bins = 50, kind = 'hist')
```

```
Out[ ]: <AxesSubplot:ylabel='Frequency'>
```



```
In [ ]: trueNewsLength['length'].describe()
```

```
Out[ ]: count    21417.000000
mean      2460.525330
std       1684.570421
min       45.000000
25%      991.000000
50%     2299.000000
75%     3314.000000
```

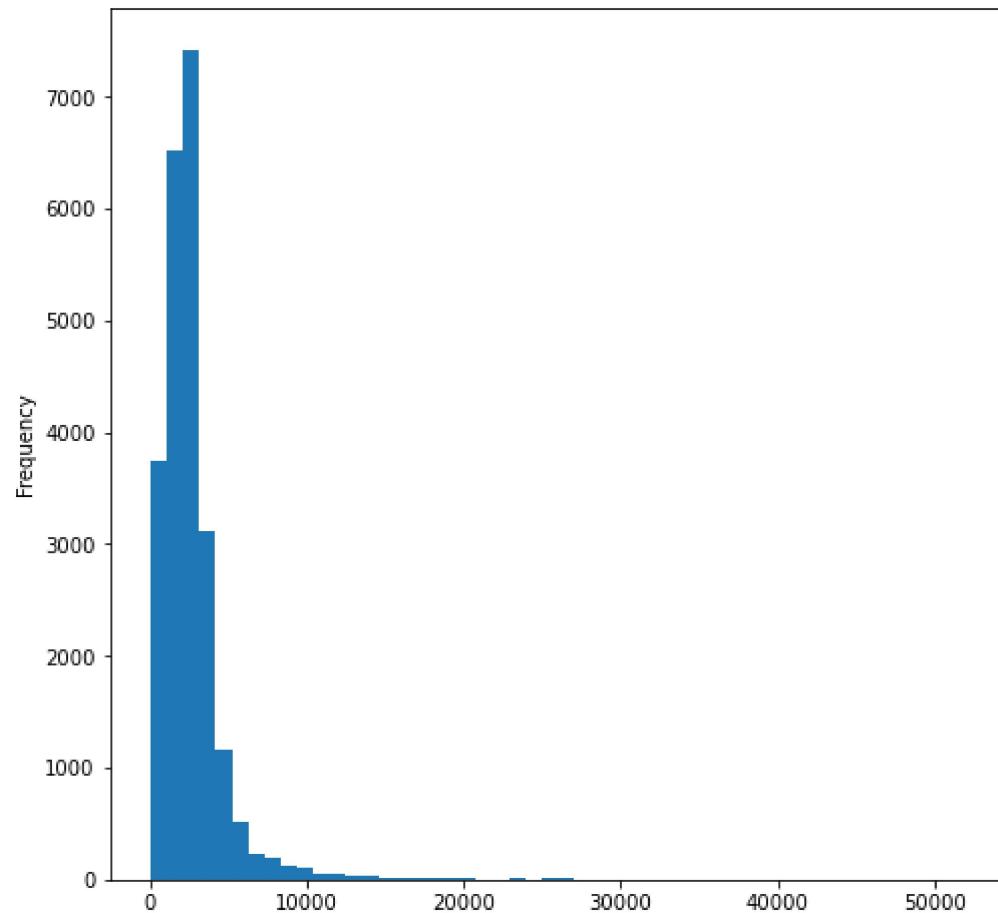
```
max      29862.000000
Name: length, dtype: float64
```

As we can see from the graph most true news articles ranged from 4500 in length to around 1000 in length with an average frequency of about 2500 articles

## Fake News

```
In [ ]: plt.figure(figsize= (8, 8))
fakeNewsLength = data.loc[data['target'] == 0, ['length']]
fakeNewsLength['length'].plot(bins = 50, kind = 'hist')
```

```
Out[ ]: <AxesSubplot:ylabel='Frequency'>
```



```
In [ ]: fakeNewsLength['length'].describe()
```

```
Out[ ]: count    23481.000000
mean     2650.776585
std      2533.693816
min      40.000000
25%     1539.000000
50%     2265.000000
75%     3133.000000
max     51905.000000
Name: length, dtype: float64
```

As we can see from the graph almost all fake news articles were around the 2500 length. This can also be seen with it's high frequency of around 7500 articles. There isn't a lot of variance in lengths

in comparison to True news articles, which have slightly broader histogram and lower frequencies.

### XIII.

TFID = TFID stands for ter frequency - inverse document frequency. It is used to give a measure to a word to show how important the word is based on the use of the word in all the documents.

Bag of words with sklearn: A bag of words in sklearn is called a CountVectorizer. the countvectorizer needs to first get imported. After it is imported it needs to be created. After that you need to feed the vectorizer documents that are in the form of a string. the data (strings) need to be fit into the vectorizer. Then after the words need to be transformed, which is when the count is created.

A TFID can be generated for the bag of words using the TfidfTransformer and TfidfVectorizer. normalization is implemented by the TFID transformer

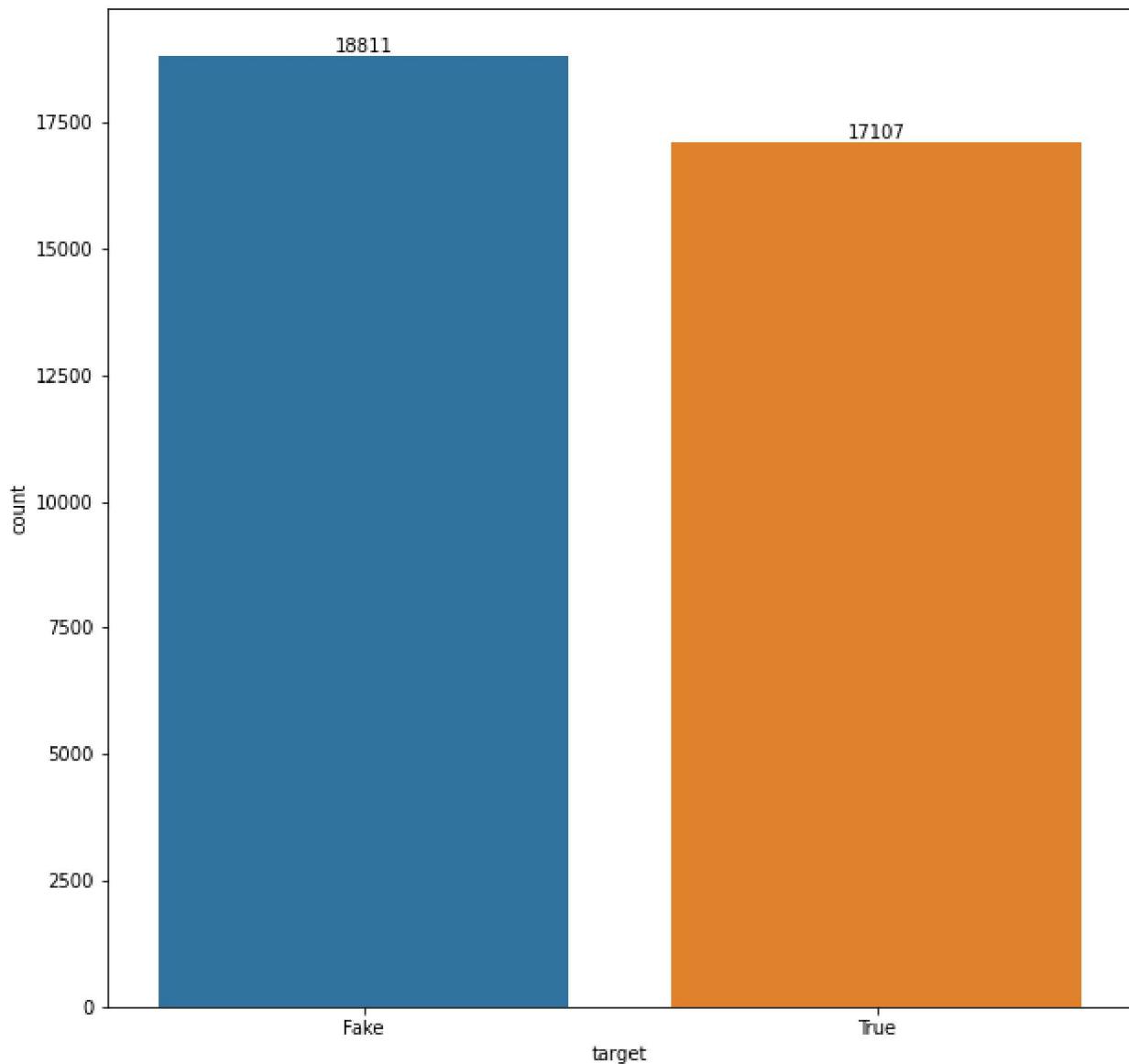
## 2. Train Test Split

```
In [ ]: AllText_train, AllText_test, target_train, target_test = train_test_split(data['AllText'])
```

### Train set count plot

```
In [ ]: plt.figure(figsize= (10, 10))
plot = sns.countplot(x = target_train)
plot.bar_label(plot.containers[0])
plot.set_xticklabels(["Fake", "True"])
plt.show
```

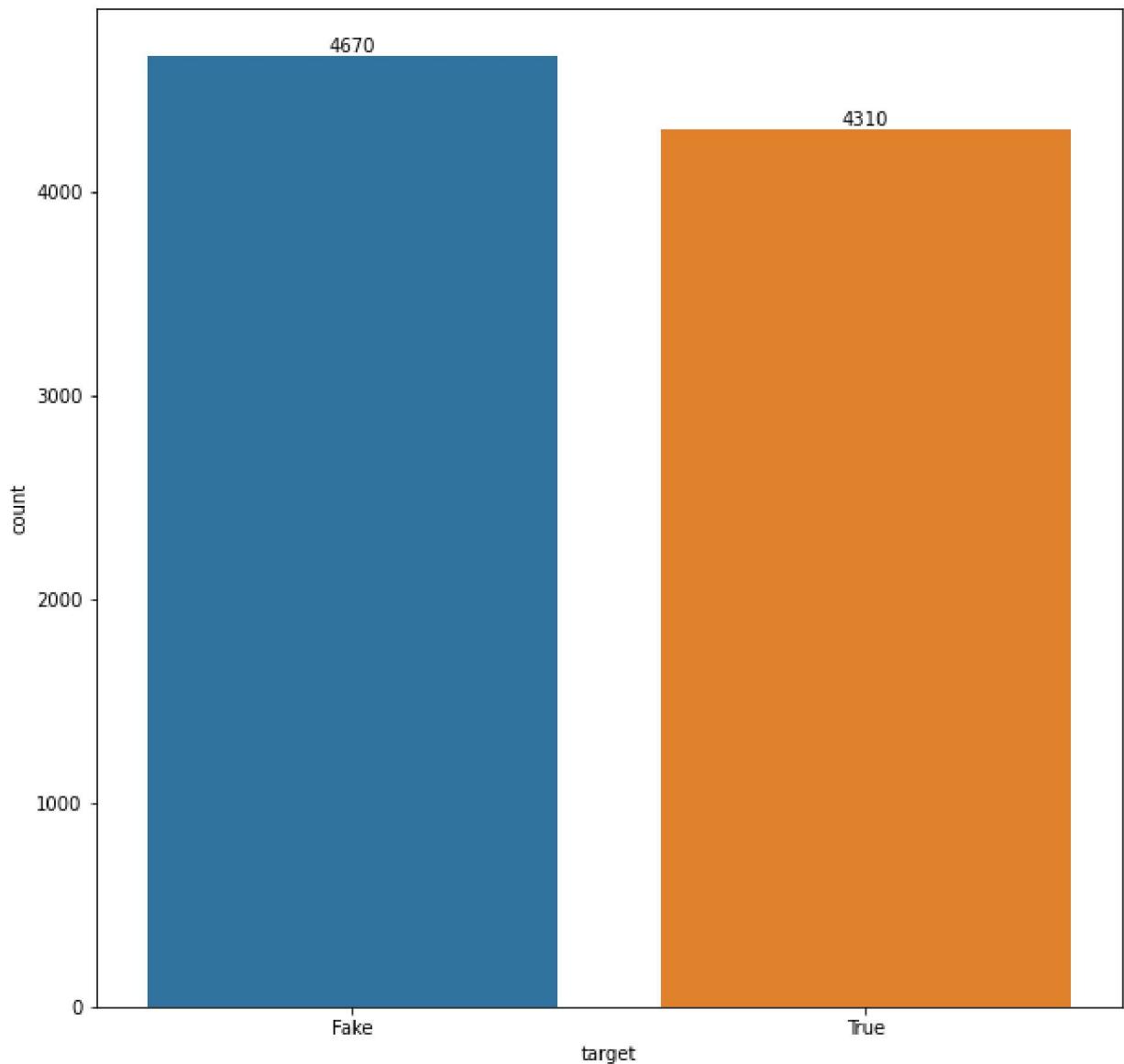
```
Out[ ]: <function matplotlib.pyplot.show(close=None, block=None)>
```



### test set countplot

```
In [ ]: plt.figure(figsize= (10, 10))
plot = sns.countplot(x = target_test)
plot.bar_label(plot.containers[0])
plot.set_xticklabels(["Fake", "True"])
plt.show

Out[ ]: <function matplotlib.pyplot.show(close=None, block=None)>
```



### 3. Training and Testing Fake News classifier using multinomialNB

```
In [ ]: stemmer = LancasterStemmer()

In [ ]: def text_process(mess):

    stopwordList = stopwords.words('english')

    # Check characters to see if they are in punctuation
    mess = [char for char in mess if char not in string.punctuation]
    # Join the characters again to form the string.
    mess = ''.join(mess)

    words = nltk.word_tokenize(mess)
    words = [t for t in words if t not in stopwordList]
    words = [stemmer.stem(w.lower()) for w in words]

    return words
```

```
In [ ]: pipelineNB = Pipeline([('bow', CountVectorizer(analyzer= text_process)), ('tfidf', TfidfTransformer()), ('classifier', MultinomialNB())])

In [ ]: pipelineNB.fit(AllText_train, target_train)

Out[ ]: Pipeline(steps=[('bow',
                         CountVectorizer(analyzer=<function text_process at 0x000001B1EBA451F0>)),
                         ('tfidf', TfidfTransformer()),
                         ('classifier', MultinomialNB())])

In [ ]: predictionMNB = pipelineNB.predict(AllText_test)

In [ ]: print(classification_report(target_test, predictionMNB))
print(confusion_matrix(target_test, predictionMNB))

precision    recall  f1-score   support

          0       0.96      0.96      0.96     4670
          1       0.96      0.95      0.95     4310

   accuracy                           0.96     8980
  macro avg       0.96      0.96      0.96     8980
weighted avg       0.96      0.96      0.96     8980

[[4478 192]
 [ 197 4113]]
```

The model is very accurate due to its high accuracy (96%). This can be seen from the classification report above the model also had a high precision score of 96%. The model does a very good job at deciding whether a news article is fake news or true news, based on the data from the classification report.

## V. news article of my choice

```
In [ ]: newArticle = "Russians plunder $5M farm vehicles from Ukraine -- to find they've been r

In [ ]: pipelineNB.predict([newArticle])

Out[ ]: array([1], dtype=int64)
```

## 4. Training and Testing a Deep Neural Network

```
In [ ]: pipelineMLP = Pipeline([('bow', CountVectorizer(analyzer=text_process)), ('tfidf', TfidfTransformer()), ('classifier', MLPClassifier())])

In [ ]: pipelineMLP.fit(AllText_train, target_train)

Iteration 1, loss = 0.21139178
```

```

Validation score: 0.994989
Iteration 2, loss = 0.01604617
Validation score: 0.996938
Iteration 3, loss = 0.00468437
Validation score: 0.997494
Iteration 4, loss = 0.00247776
Validation score: 0.997773
Iteration 5, loss = 0.00171587
Validation score: 0.997773
Iteration 6, loss = 0.00134731
Validation score: 0.997773
Iteration 7, loss = 0.00113292
Validation score: 0.997773
Iteration 8, loss = 0.00099114
Validation score: 0.997773
Iteration 9, loss = 0.00089001
Validation score: 0.997773
Iteration 10, loss = 0.00081235
Validation score: 0.998051
Iteration 11, loss = 0.00075018
Validation score: 0.998051
Iteration 12, loss = 0.00069816
Validation score: 0.998051
Iteration 13, loss = 0.00065360
Validation score: 0.997773
Iteration 14, loss = 0.00061427
Validation score: 0.998051
Iteration 15, loss = 0.00057921
Validation score: 0.998051
Iteration 16, loss = 0.00054732
Validation score: 0.998051
Iteration 17, loss = 0.00051819
Validation score: 0.998051
Iteration 18, loss = 0.00049129
Validation score: 0.997773
Iteration 19, loss = 0.00046628
Validation score: 0.997773
Iteration 20, loss = 0.00044299
Validation score: 0.997773
Iteration 21, loss = 0.00042098
Validation score: 0.997773
Validation score did not improve more than tol=0.000100 for 10 consecutive epochs. Stopping.

```

```

Out[ ]: Pipeline(steps=[('bow',
                         CountVectorizer(analyzer=<function text_process at 0x000001B1EBA451F0>)),
                         ('tfidf', TfidfTransformer()),
                         ('classifier',
                          MLPClassifier(early_stopping=True, hidden_layer_sizes=(100, 4),
                                         random_state=0, verbose=2))])

```

```
In [ ]: predictionMLP = pipelineMLP.predict(AllText_test)
```

```
In [ ]: print(classification_report(target_test, predictionMLP))
print(confusion_matrix(target_test, predictionMLP))
```

precision	recall	f1-score	support

	0	1.00	0.99	1.00	4670
	1	0.99	1.00	1.00	4310
accuracy				1.00	8980
macro avg		1.00	1.00	1.00	8980
weighted avg		1.00	1.00	1.00	8980

```
[[4646  24]
 [ 11 4299]]
```

As we can see from the confusion matrix and classification report, the deep neural network does a lot better job a classification with a 100% accuracy rate. The model does a really good job at predicting true news from fake news with almost no missclassifications. The downside would be the time taken to fit the model. It considerably longer than MultinomialNB, but the increase in accuracy makes it worth it. Its precision also dips to 99%, but that should not be too much of a problem. Since we can see from the the confusion matrix a very little amount was actually missclassified.

## V.Deep neural network model on my news article

```
In [ ]: pipelineMLP.predict([newArticle])
```

```
Out[ ]: array([0], dtype=int64)
```

## VII. discussion

There is a somewhat noticeable difference between the performance of the NB model and the deep learning model. The most obvious one can be seen by how it predicts on the news article I found. The NB model predicts a 1 meaning it is true news, while the MLP (deep learning model) predicts a 0 meaning fake news. The MLP model also has an accuracy of 100% in comparison to the 96% accuracy of the NB model. The MLP model has a higher success rate than the NB model. the MLP model also averages a value of 1.0 or 100% for most of the data in the classification matrix, while the NB averages a value of .96 or 96%. Further reinforcing the fact that the MLP (deep neural network) is far more accurate than the NB model.

## 5. Extra credit

```
In [ ]: data_no_alltext = data_copy.drop(['AllText', 'target', 'date'], axis= 1)
```

```
In [ ]: data_no_alltext
```

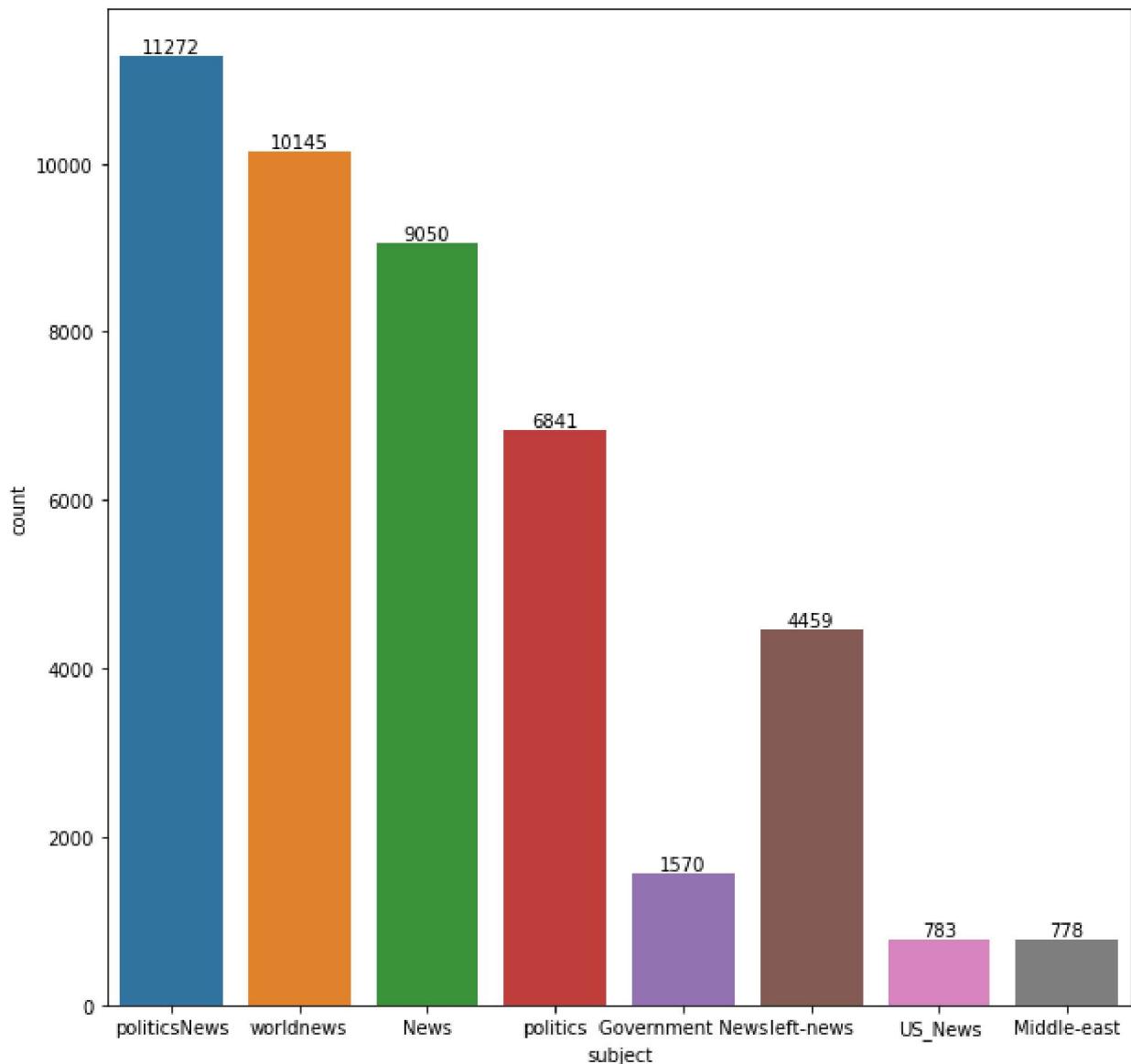
	title	text	subject
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews

		title	text	subject
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...		politicsNews
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donald...		politicsNews
...	...	...	...	...
44893	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...		Middle-east
44894	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It's a familiar theme. ...		Middle-east
44895	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...		Middle-east
44896	How to Blow \$700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...		Middle-east
44897	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...		Middle-east

44898 rows × 3 columns

```
In [ ]: plt.figure(figsize= (10, 10))
plot = sns.countplot(x = 'subject', data= data_copy)
plot.bar_label(plot.containers[0])
plt.show
```

```
Out[ ]: <function matplotlib.pyplot.show(close=None, block=None)>
```



```
In [ ]: cols = ['title', 'text']
data_no_alltext['AllText'] = data_no_alltext[cols].apply(lambda row: ' '.join(row.value
```

```
In [ ]: data_no_alltext
```

	title	text	subject	AllText
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	As U.S. budget fight looms, Republicans flip t...
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	U.S. military to accept transgender recruits o...
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	Senior U.S. Republican senator: 'Let Mr. Muell...
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	FBI Russia probe helped by Australian diplomat...

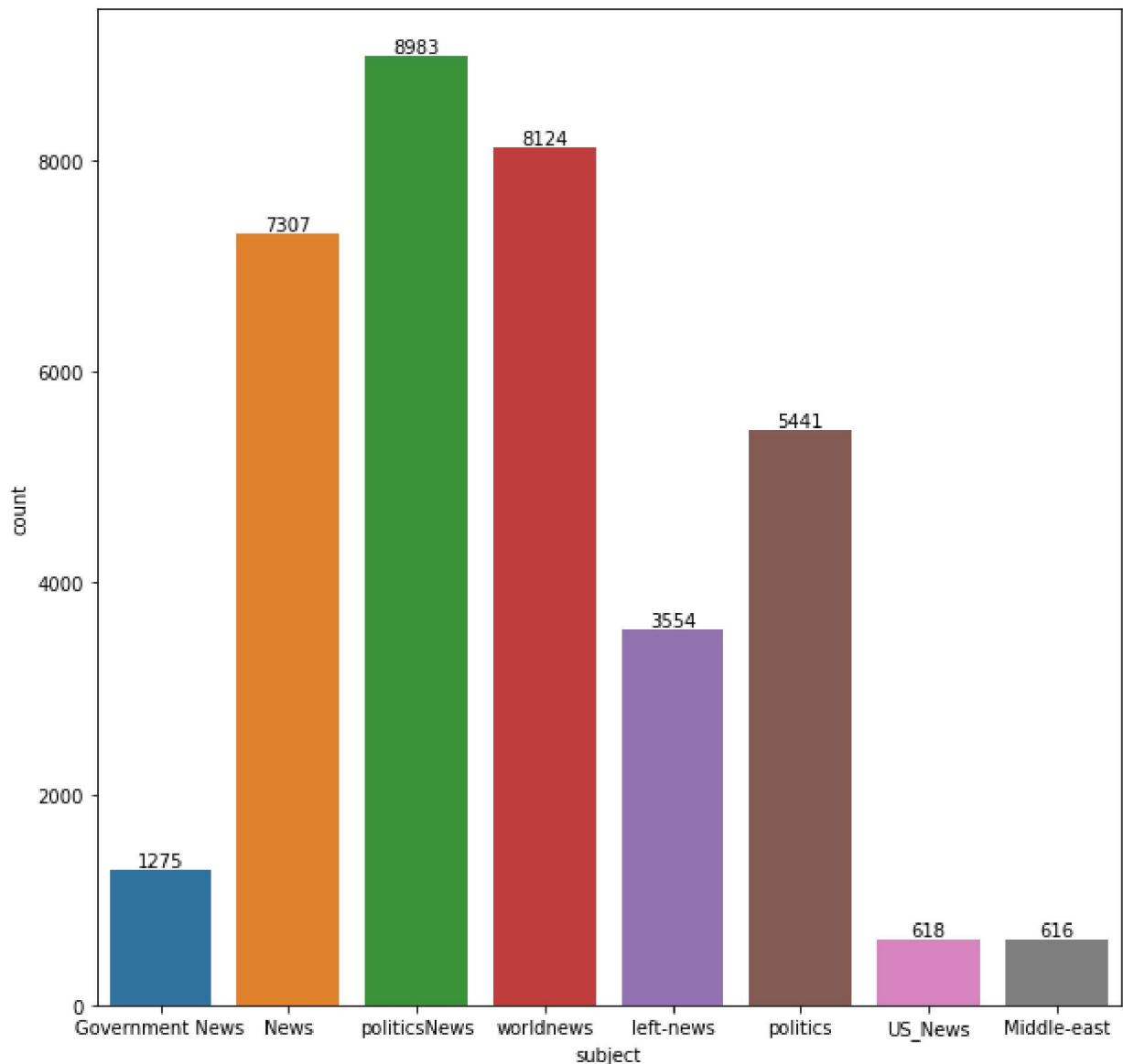
	<b>title</b>	<b>text</b>	<b>subject</b>	<b>AllText</b>
4	Trump wants Postal Service to charge 'much more'	SEATTLE/WASHINGTON (Reuters) - President Donald Trump wants Postal Service to charge 'much more'	politicsNews	Trump wants Postal Service to charge 'much more'
...	...	...	...	...
44893	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	McPain: John McCain Furious That Iran Treated ...
44894	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It's a familiar theme. ...	Middle-east	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...
44895	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	Sunnistan: US and Allied 'Safe Zone' Plan to T...
44896	How to Blow \$700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	How to Blow \$700 Million: Al Jazeera America F...
44897	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	10 U.S. Navy Sailors Held by Iranian Military ...
				...

44898 rows × 4 columns

```
In [ ]: alltex_train, alltex_test, subject_train, subject_test = train_test_split(data_no_allte)
```

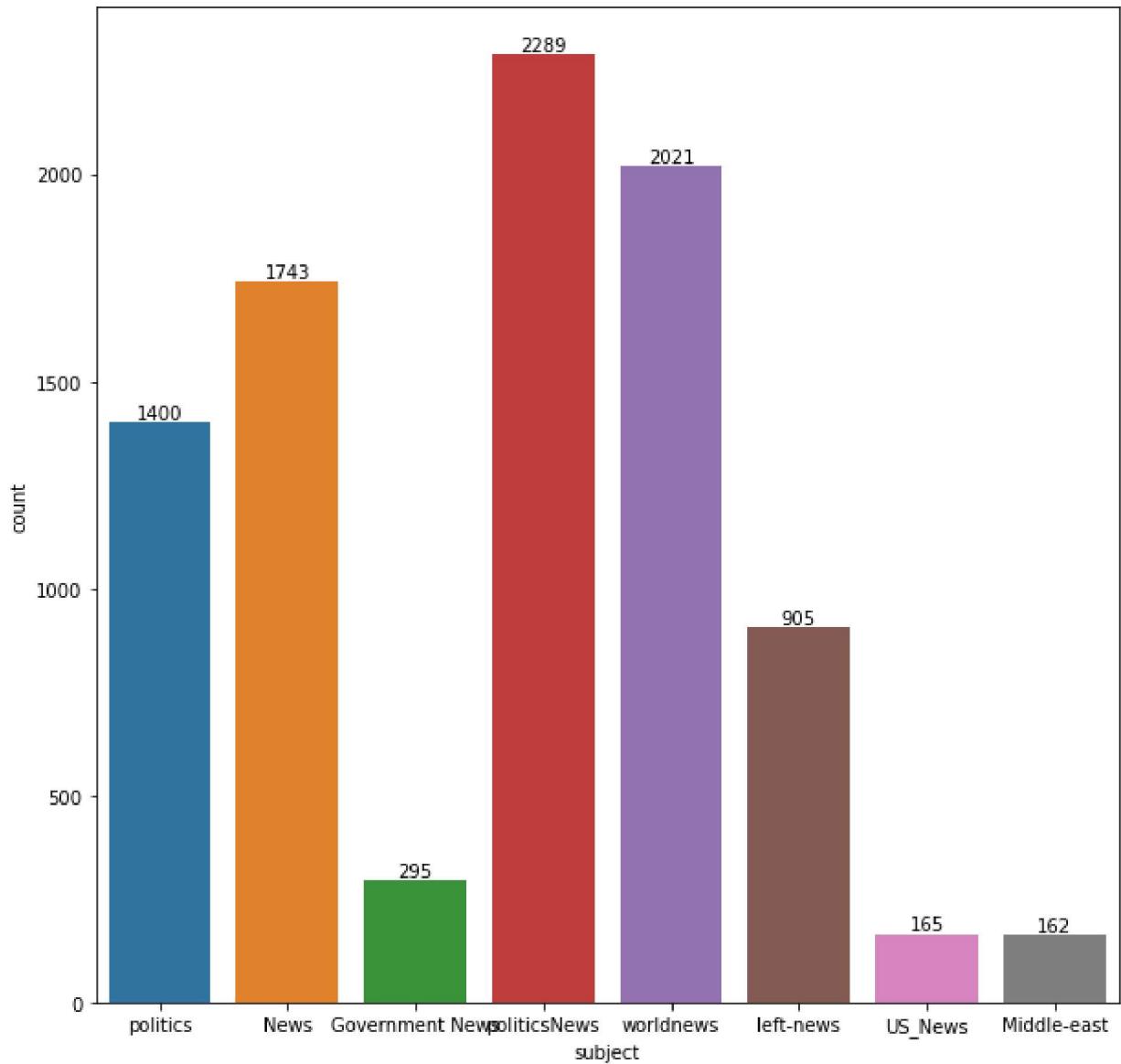
```
In [ ]: plt.figure(figsize= (10, 10))
plot = sns.countplot(x = subject_train)
plot.bar_label(plot.containers[0])
plt.show
```

```
Out[ ]: <function matplotlib.pyplot.show(close=None, block=None)>
```



```
In [ ]: plt.figure(figsize= (10, 10))
plot = sns.countplot(x = subject_test)
plot.bar_label(plot.containers[0])
plt.show
```

```
Out[ ]: <function matplotlib.pyplot.show(close=None, block=None)>
```



```
In [ ]: pipelineSubject = Pipeline([('bow', CountVectorizer(analyzer=text_process)), ('tfidf',
```

```
In [ ]: pipelineSubject.fit(alltex_train, subject_train)
```

Iteration 1, loss = 1.21053643  
 Validation score: 0.778396  
 Iteration 2, loss = 0.45438485  
 Validation score: 0.812361  
 Iteration 3, loss = 0.31417595  
 Validation score: 0.793987  
 Iteration 4, loss = 0.26565381  
 Validation score: 0.767539  
 Iteration 5, loss = 0.23382729  
 Validation score: 0.756682  
 Iteration 6, loss = 0.21471861  
 Validation score: 0.750278  
 Iteration 7, loss = 0.20241891  
 Validation score: 0.748886  
 Iteration 8, loss = 0.19450051  
 Validation score: 0.747216

Validation score did not improve more than tol=0.000100 for 5 consecutive epochs. Stopping.

```
Out[ ]: Pipeline(steps=[('bow',
                         CountVectorizer(analyzer=<function text_process at 0x000001B1EBA451F0
                       >)),
                     ('tfidf', TfidfTransformer()),
                     ('classifier',
                      MLPClassifier(early_stopping=True,
                                    hidden_layer_sizes=(100, 50, 8), max_iter=22,
                                    n_iter_no_change=5, random_state=0,
                                    verbose=2))])
```

```
In [ ]: predictionSubject = pipelineSubject.predict(alltex_test)
```

```
In [ ]: print(classification_report(subject_test, predictionSubject))
print(confusion_matrix(subject_test, predictionSubject))
```

	precision	recall	f1-score	support
Government News	0.20	0.04	0.07	295
Middle-east	0.00	0.00	0.00	162
News	0.95	0.98	0.96	1743
US_News	0.50	0.87	0.64	165
left-news	0.14	0.02	0.03	905
politics	0.51	0.87	0.65	1400
politicsNews	0.97	0.98	0.97	2289
worldnews	0.96	0.98	0.97	2021
accuracy			0.81	8980
macro avg	0.53	0.59	0.54	8980
weighted avg	0.76	0.81	0.77	8980

```
[[ 13   0    8   0    3  255    7   9]
 [ 6   0    4  141   0    0   11   0]
 [ 0   0  1707   0    2   30    0   4]
 [ 5   2    6  143   0    2    3   4]
 [ 11   0   20   0   14  854    2   4]
 [ 30   0   50   0   76 1217   17  10]
 [ 0   0    1   0    2   10 2232   44]
 [ 0   0    0   0    0    4   38 1979]]
```

From what we can see the models accuracy is 81%, this is shown within the classification report. But there is a major problem, since most of data is coming from politicalNews and worldNews, so it is somewhat unbalanced when classifying. Another interesting thing within the classification matrix is that Middle-east has 0 for precision recall and f1-score, this could be an issue with how the model is classifying the data. The model is good at classifying politicsnews and world news along with news, but it has a hard time with government news and left news. The macro avg are also somewhat low since they are close to the 50% mark