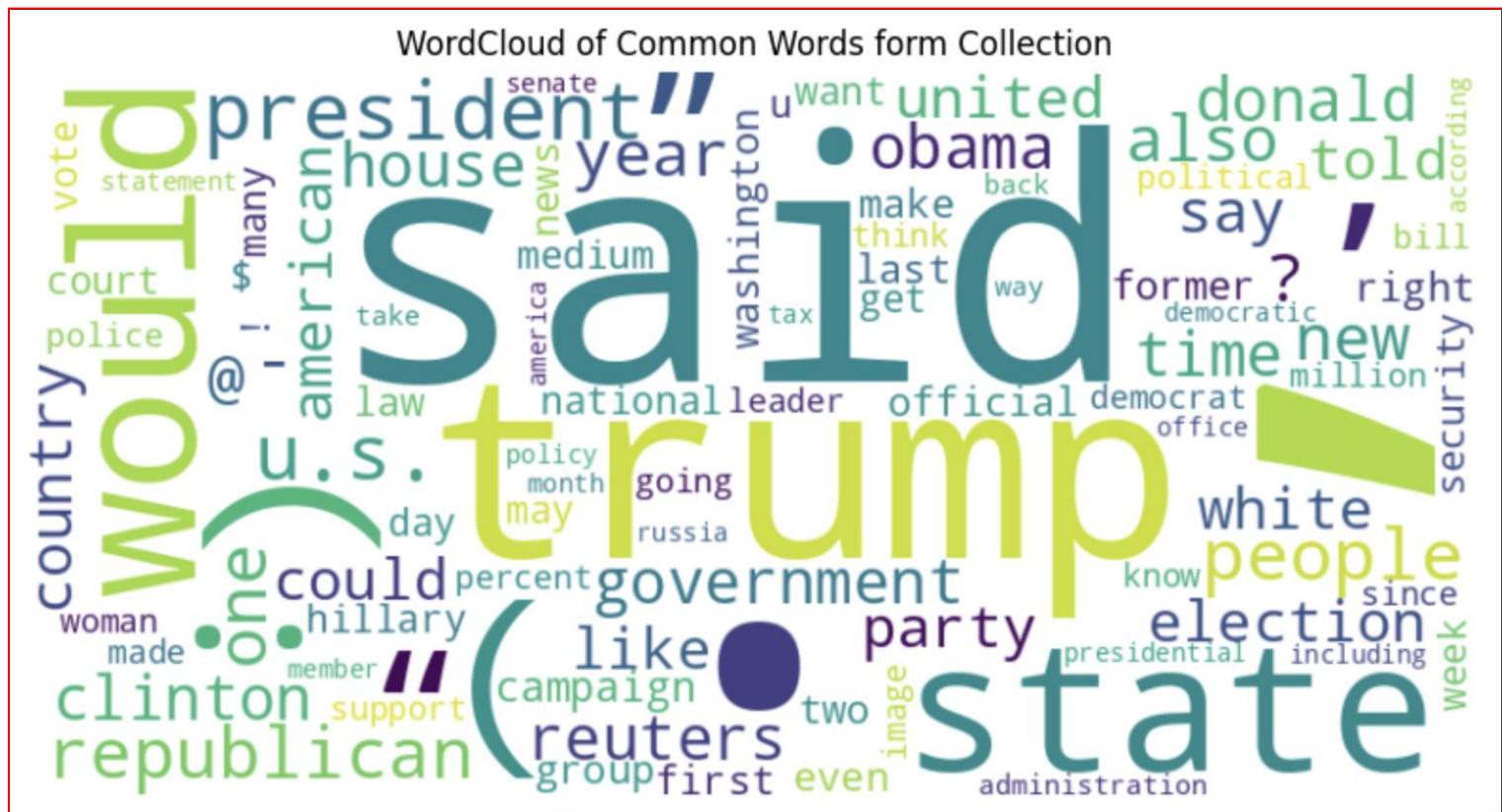


Homework Assignment 1: Fake News Detection via Text Classification

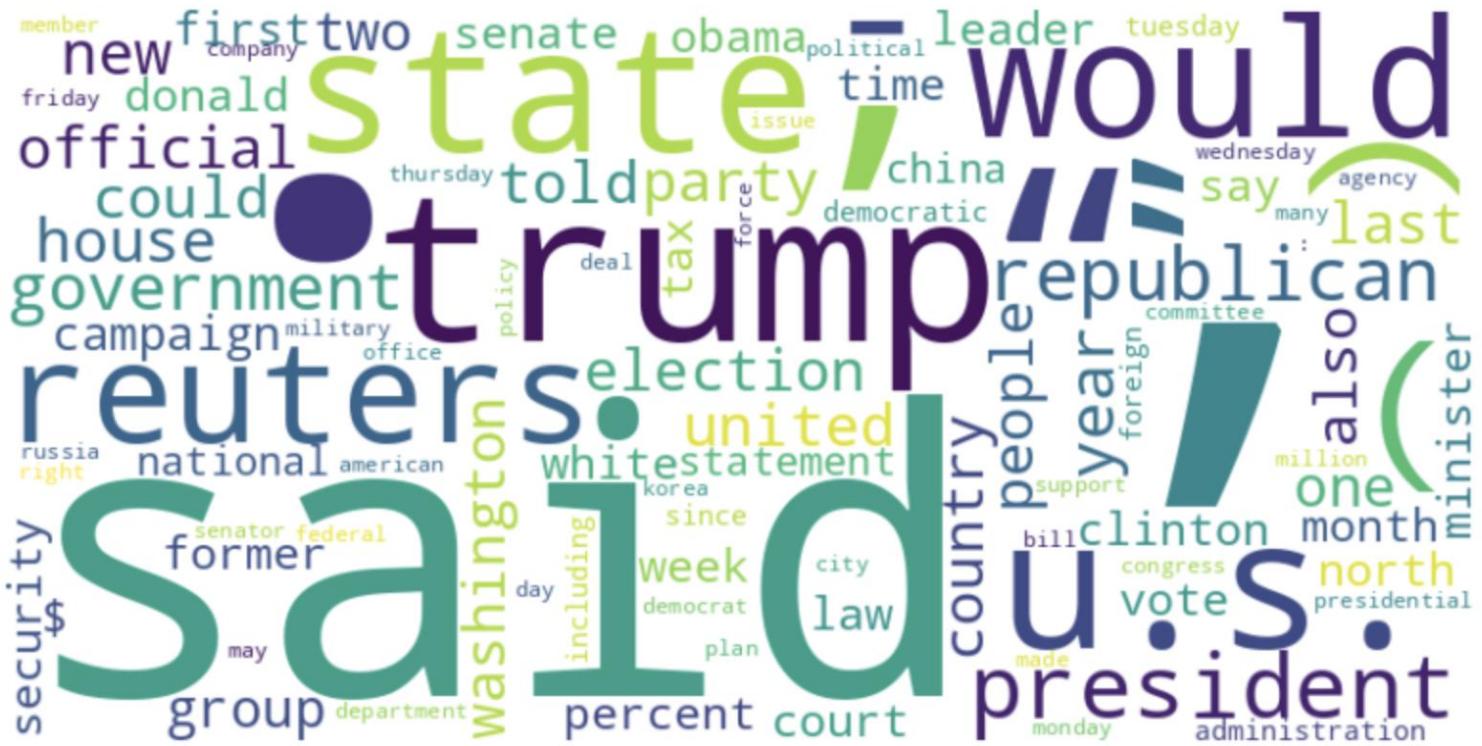
Task 1 - Explore Essential Information from Text Data and Preprocessing

In this dataset, you will have access to both real news and fake news. Before machine learning, please explore the essential information from the textual data, e.g., the most commonly used words in the collection/real news/fake news. Note that you need to call “stopword removal” and “lemmatization” functions before calculating the word frequency. Then, please answer the following questions:

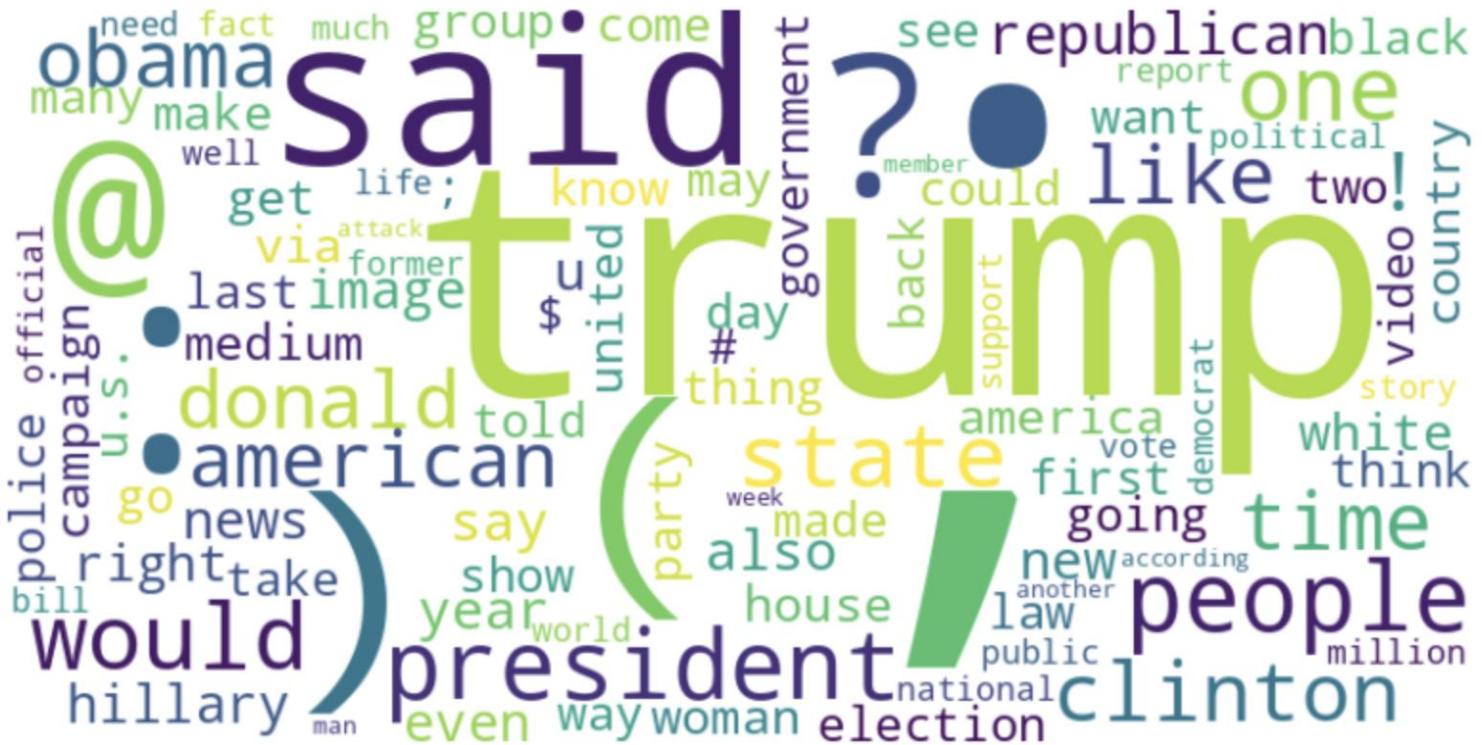
1. What are the most commonly used words (top 100) in the collection, the most commonly used words (top 100) in the real news and most commonly used words (top 100) in the fake news?



WordCloud of Common Words from Real News



WordCloud of Common Words form Fake News



2. By reading the preprocessed textual data, can you easily tell the difference between the real news and fake news? What does the strongest feature set (for machine learning) look like?

Before pre-processing the text

```
1 true_df['all_text'].iloc[0]
```

'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 the 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.' □'

```
1 fake_df['all_text'].iloc[0]
```

' Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing Donald Trump just couldn't wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn't do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year, President Angry Pants tweeted. 2018 will be a great year for America! As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year. 2018 will be a great year for America! Donald J. Trump (@realDonaldTrump) December 31, 2017Trump's tweet went down about as well as you'd expect.What kind of president sends a New Year's greeting like this despicable, petty, infantile gibberish? Only Trump! His lack of decency won't even allow him to rise above the gutter long enough to wish the American citizens a happy new year! Bishop Talbert Swan (@talbertswan) December 31, 2017No one likes you Calvin (@calvinstowell) December 31, 2017Your impeachment would make 2018 a great year for America, but I'll also accept regaining control of Congress. Miranda Yaver (@mirandayaver) December 31, 2017Do you hear yourself talk? When you have to include that many people that hate you you have to wonder? Why do they all hate me? Alan Sandoval (@AlanSandoval13) December 31, 2017Who uses the word Haters in a New Year's wish?? Marlene (@marlene399) December 31, 2017You can't just say happy new year? Karen Pollitt (@korenccarpenter) December 31, 2017Here's Trump's New Year's Eve tweet from 2016.Happy New Year to all, including to my many enemies and those who have fought me and lost so badly they just don't know what to do. Love! Donald J. Trump (@realDonaldTrump) December 31, 2016This is nothing new for Trump. He's been doing this for years.Trump has directed messages to his enemies and haters for New Year's, Easter, Thanksgiving, and the anniversary of 9/11. pic.twitter.com/4FPAe2KypA Daniel Dale (@ddale8) December 31, 2017Trump's holiday tweets are clearly not presidential.How long did he work at Hallmark before becoming President? Steven Goodine (@SGGoodine) December 31, 2017He's always been like this... . the only difference is that in the last few years, his filter has been breaking down. Roy Schulze (@thbhthtt) December 31, 2017Who, apart from a teenager uses the term haters? Wendy (@WendyWhistles) December 31, 2017he's a fucking 5 year old Who Knows (@rainyday80) December 31, 2017So, to all the people who voted for this a hole thinking he would change once he got into power, you were wrong! 70-year-old men don't change and now he's a year older.Photo by Andrew Burton/Getty Images.'

After pre-processing the text

```
1 true_df['all_text'].iloc[0]
```

'u.s. budget fight loom , republican flip fiscal script washington (reuters) - head conservative republican faction u.s. congress , voted month huge expansion national debt pay tax cut , called " fiscal conservative " sunday urged budget restraint 2018. keeping sharp pivot way among republican , u.s. representative mark meadow , speaking cbs ' " face nation , " drew hard line federal spending , lawmaker bracing battle january . return holiday wednesday , lawmaker begin trying pas federal budget fight likely linked issue , immigration policy , even november congressional election campaign approach republican seek keep control congress . president donald trump republican want big budget increase military spending , democrat also want proportional increase non-defense " discretionary " spending program support education , scientific research , infrastructure , public health environmental protection . " (trump) administration already willing say : ' ' going increase non-defense discretionary spending ... 7 percent , " meadow , chairman small influential house freedom caucus , said program . " , democrat saying ' enough , need give government pay raise 10 11 percent . fiscal conservative , ' see rationale . . . eventually run people ' money , " said . meadow among republican voted late december party ' debt-financed tax overhaul , expected balloon federal budget deficit add \$ 1.5 trillion 10 year \$ 20 trillion national debt . " ' interesting hear mark talk fiscal responsibility , " democratic u.s. representative joseph crowley said cbs . crowley said republican tax bill would require united states borrow \$ 1.5 trillion , paid future generation , finance tax cut corporation rich . " one least ... fiscally responsible bill ' ever seen passed history house representative . think ' going paying many , many year come , " crowley said . republican insist tax package , biggest u.s. tax overhaul 30 year , boost economy job growth . house speaker paul ryan , also supported tax bill , recently went meadow , making clear radio interview welfare " entitlement reform , " party often call , would top republican priority 2018. republican parlance , " entitlement " program mean food stamp , housing assistance , medicare medicaid health insurance elderly , poor disabled , well program created washington assist needy . democrat seized ryan ' early december remark , saying showed republican would try pay tax overhaul seeking spending cut social program . goal house republican may take back seat senate , vote democrat needed approve budget prevent government shutdown . democrat use leverage senate , republican narrowly control , defend discretionary non-defense program social spending , tackling issue " dreamer , " people brought illegally country child . trump september put march 2018 expiration date deferred action childhood arrival , daca , program , protects young immigrant deportation provides work permit . president said recent twitter message want funding proposed mexican border wall immigration law change exchange agreeing help dreamer . representative debbie dingell told cbs favor linking issue policy objective , wall funding . " need daca clean , " said . wednesday , trump aide meet congressional leader discuss issue . followed weekend strategy session trump republican leader jan. 6 7 , white house said . trump also scheduled meet sunday florida republican governor rick scott , want emergency aid . house passed \$ 81 billion aid package hurricane florida , texas puerto rico , wildfire california . package far exceeded \$ 44 billion requested trump administration . senate yet voted aid .<'

```
1 fake_df['all_text'].iloc[0]
```

'donald trump sends embarrassing new year ' eve message ; disturbing donald trump wish american happy new year leave . instead , give shout enemy , hater dishonest fake news medium . former reality show star one job . country rapidly grows stronger smarter , want wish friend , supporter , enemy , hater , even dishonest fake news medium , happy healthy new year , president angry pant tweeted . 2018 great year america ! country rapidly grows stronger smarter , want wish friend , supporter , enemy , hater , even dishonest fake news medium , happy healthy new year . 2018 great year america ! donald j. trump (@realDonaldTrump) december 31 , 2017trump tweet went well expect.what kind president sends new year greeting like despicable , petty , infantile gibberish ? trump ! lack decency even allow rise gutter long enough wish american citizen happy new year ! bishop talbert swan (@talbertswan) december 31 , 2017no one like calvin (@calvinstowell) december 31 , 2017your impeachment would make 2018 great year america , also accept regaining control congress . miranda yaver (@mirandayaver) december 31 , 2017do hear talk ? include many people hate wonder ? hate ? alan sandoval (@alansandoval13) december 31 , 2017who us word hater new year wish ? marlene (@marlene399) december 31 , 2017you say happy new year ? karen pollitt (@korenccarpenter) december 31 , 2017here trump new year eve tweet 2016.happy new year , including many enemy fought lost badly know . love ! donald j. trump (@realDonaldTrump) december 31 , 2016this nothing new trump . years.trump directed message enemy hater new year , easter , thanksgiving , anniversary 9/11 . pic.twitter.com/4FPAe2KypA daniel dale (@ddale8) december 31 , 2017trump holiday tweet clearly presidential.how long work hallmark becoming president ? steven goodine (@sggoodine) december 31 , 2017he always like . . . difference last year , filter breaking . roy schulze (@thbhthtt) december 31 , 2017who , apart teenager us term hater ? wendy (@wendywhistles) december 31 , 2017the fucking 5 year old know (@rainyday80) december 31 , 2017so , people voted hole thinking would change got power , wrong ! 70-year-old men change year old.photo andrew burton/getty image .<'

Real News Articles

- The language is formal and professional and proper grammar.
- The content focuses on political events, budget discussions, and legislative priorities.

- c. Statements are attributed to specific individuals and politicians.
- d. The tone is serious and focused on policy matters.
- e. The article discusses budgetary issues, government policies, and political strategies.
- f. Typically focus on factual information.

Fake News Articles

- a. The language is informal language, filled with grammatical errors, and abusive vocabulary.
- b. The content focuses on personal attacks, opinions, and criticisms of public figures.
- c. Statements are made without specific attribution or sources.
- d. The tone is sensationalist and derogatory towards the subject, and sometimes a sensationalist or exaggerated tone.
- e. The article contains personal opinions, criticisms, and insults directed towards political figures.
- f. Typically focus on conspiracy theories or biased interpretations of events.

Strongest Feature Set for Machine Learning

- a. For distinguishing between real and fake news based on text content, features related to language formality, subject matter, attribution of statements, and tone could be essential.
- b. Language complexity metrics (e.g., average word length, sentence length, vocabulary richness) may differentiate real news from fake news.
- c. The presence of factual information versus opinionated or speculative content could also be a crucial feature.
- d. Additionally, sentiment analysis to identify positive or negative tones and the presence of specific keywords or phrases associated with either real or fake news could strengthen the feature set for machine learning classification.
- e. Well-known and real publishing platforms and historical data on the reliability of news sources could further enhance the model's performance.

EDA before data preprocessing

```
[8] 1 fake_df.shape,true_df.shape
```

```
((23481, 4), (21417, 4))
```

```
▶ 1 fake_df.isnull().sum(),true_df.isnull().sum()
```

```
✉ (title      0  
text       0  
subject    0  
date       0  
dtype: int64,  
title      0  
text       0  
subject    0  
date       0  
dtype: int64)
```

```
[ ] 1 fake_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 23481 entries, 0 to 23480  
Data columns (total 4 columns):  
 # Column Non-Null Count Dtype  
---  
 0 title    23481 non-null object  
 1 text     23481 non-null object  
 2 subject  23481 non-null object  
 3 date     23481 non-null object  
 dtypes: object(4)  
 memory usage: 733.9+ KB
```

```
▶ 1 all_news_df=pd.concat([fake_df,true_df],ignore_index=True)  
2 all_news_df.shape
```

```
✉ (44689, 4)
```

```
[ ] 1 all_news_df['subject'].value_counts()
```

```
politicsNews    11220  
worldnews      9991  
News            9050  
politics        6838  
left-news       4459  
Government News 1570  
US_News          783  
Middle-east      778  
Name: subject, dtype: int64
```

```
[ ] 1 fake_df['subject'].value_counts()
```

```
News            9050  
politics        6838  
left-news       4459  
Government News 1570  
US_News          783  
Middle-east      778  
Name: subject, dtype: int64
```

```
[ ] 1 true_df['subject'].value_counts()
```

```
politicsNews    11220  
worldnews      9991  
Name: subject, dtype: int64
```

```
[ ] 1 true_df.head(2)
```

subject	label	all_text
0 politicsNews	1	As U.S. budget fight looms, Republicans flip t...
1 politicsNews	1	U.S. military to accept transgender recruits o...

```
[ ] 1 true_df['all_text'].iloc[0]
```

```
'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-...'
```

```
[ ] 1 fake_df['all_text'].iloc[0]
```

```
' Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing Donald Trump just couldn't wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn't do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year, President Angry Pants reelected. 2018 will be a great year for America! As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year. 2018 will be a great year for America! Donald J. Trump (@realDonaldTrump) December 31, 2017Trump's tweet went down about as well as you'd expect.What kind of president sends a New Year's greeting like this d...'
```

```
[ ] 1 true_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 21417 entries, 0 to 21416  
Data columns (total 4 columns):  
 # Column Non-Null Count Dtype  
---  
 0 title    21417 non-null object  
 1 text     21417 non-null object  
 2 subject  21417 non-null object  
 3 date     21417 non-null object  
 dtypes: object(4)  
memory usage: 669.4+ KB
```

```
▶ 1 fake_df[fake_df.duplicated()].shape,true_df[true_df.duplicated()].shape
```

```
✉ ((3, 4), (206, 4))
```

```
[9] 1 fake_df=fake_df.drop_duplicates()  
2 true_df=true_df.drop_duplicates()  
3  
4 print(fake_df.shape,true_df.shape)  
5 print(fake_df[fake_df.duplicated()].shape,true_df[true_df.duplicated()].shape)
```

```
(23478, 4) (21211, 4)
```

```
(0, 4) (0, 4)
```

```
[10] 1 fake_df=fake_df.drop(columns=['date'])  
2 true_df=true_df.drop(columns=['date'])
```

```
[11] 1 fake_df['title']=fake_df['title'].astype(str)  
2 fake_df['text']=fake_df['text'].astype(str)  
3 true_df['title']=true_df['title'].astype(str)  
4 true_df['text']=true_df['text'].astype(str)
```

```
[12] 1 fake_df['label']=0  
2 true_df['label']=1
```

```
1 true_df['all_text']=true_df['title']+ ' '+true_df['text']  
2 fake_df['all_text']=fake_df['title']+ ' '+fake_df['text']  
3 all_news_df['all_text']=all_news_df['title']+ ' '+all_news_df['text']  
4  
5 true_df=true_df[['subject','label','all_text']]  
6 fake_df=fake_df[['subject','label','all_text']]  
7 all_news_df=all_news_df[['subject','label','all_text']]
```

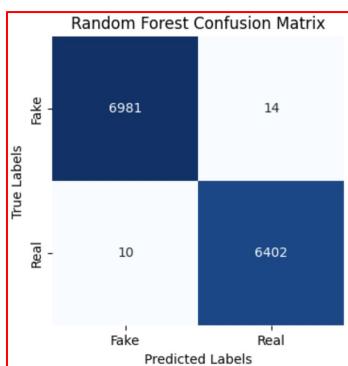
Task 2 - Build Machine Learning Model

In this task, please build machine learning text classification models to classify all the news into "real" and "fake" categories. You will need to split the data into training and testing collections for this task (e.g., 70% and 30%). Note that you can choose different kinds of algorithms, e.g., regression and MultinomialNB, and you can feed machine learning algorithms different kinds of feature sets, e.g., "term-frequency" and "TFIDF" as the feature set.

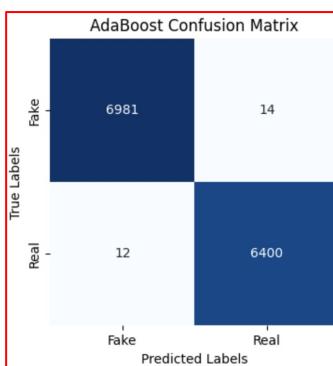
	ML Model	Feature	Precision	Recall	Accuracy
31	<u>Random Forest</u>	TF-IDF (Term Frequency & Inverse Document Frequency)	0.998129	0.998284	0.998284
32	Random Forest	Count Vectorizer (Bag of words method)	0.997818	0.998596	0.998284
30	Random Forest	Term Frequency (TF – without IDF)	0.997818	0.998284	0.998135
35	<u>Adaptive Boosting</u>	Term Frequency (TF – without IDF)	0.997817	0.998129	0.998061
37	Adaptive Boosting	Count Vectorizer (Bag of words method)	0.997507	0.998596	0.998135
40	Gradient Boosting	Term Frequency (TF – without IDF)	0.997351	0.998284	0.997912
36	Adaptive Boosting	TF-IDF (Term Frequency & Inverse Document Frequency)	0.997197	0.998596	0.997986
7	<u>Logistic Regression</u>	Count Vectorizer (Bag of words method)	0.996723	0.996101	0.996569
41	Gradient Boosting	TF-IDF (Term Frequency & Inverse Document Frequency)	0.995955	0.99844	0.997315
42	Gradient Boosting	Count Vectorizer (Bag of words method)	0.995955	0.99844	0.997315
15	Passive Aggressive Classifier	Term Frequency (TF – without IDF)	0.995946	0.996101	0.996196
27	Stochastic Gradient Descent	Count Vectorizer (Bag of words method)	0.995463	0.992358	0.994182

Random Forest gave the best results for the TF-IDF and Bag of Words method with Adaptive Boosting being the second using only the TF and Bag of Words method. Since both Random Forest and AdaBoost are tree-based and combine weak learners, we will delve deeper with three models namely Random Forest with TF-IDF, AdaBoost with TF and Logistic Regression with Bag of Words method.

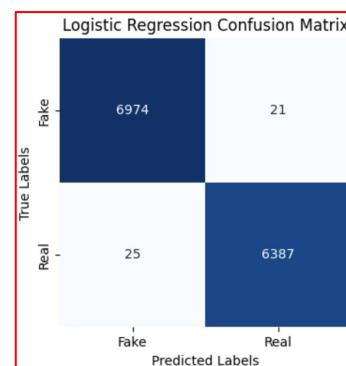
Random Forest with TF-IDF



AdaBoost with TF



Logistic Regression with Bag of Words



We tried a total of 45 (9 models * 5 different features) models out of which the Random Forest and AdaBoost models appear to have similar outcomes, while the Logistic Regression model has slightly more misclassifications. Random Forest and AdaBoost were very close in prediction where each predicted only 14 instances to be Real when it was coming from the fake source. Adaboost missed 2 additional data points in predicting as Real compared to Random Forest. Logistic Regression being a very simple model in comparison to Random Forest and Adaboost still performed a lot better and was able to capture the nuances and complex patterns in the dataset. It was off by an extra 7 records when predicting the data points as fake when compared to both the other models. Logistic Regression is a linear model that makes predictions based on a linear combination of input features. Random Forest is a more complex model capable of capturing non-linear relationships. Random Forest has a higher tendency to overfit compared to Logistic Regression, especially when the dataset is noisy or contains irrelevant features.

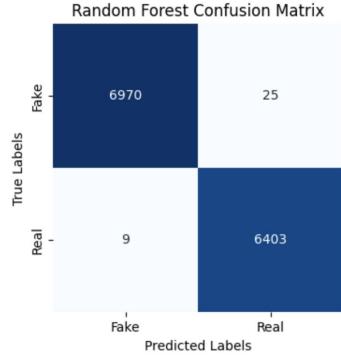
Task 3 – Enhanced NLP Features

After applying "POS Tagging", you can locate specific kinds of words in the collection, e.g., nouns and verbs. Please build additional classifier(s) to classify the news by leveraging POS information, e.g., only use the "nouns" or "adj" + "noun" as features. Did you witness the performance improvement (compared with the result from task 2)? why

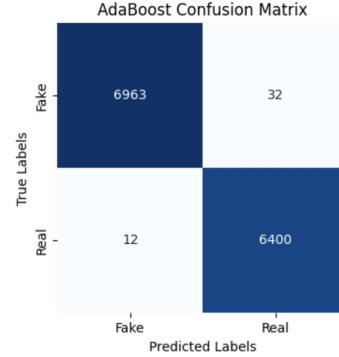
	ML Model	Feature	Filter	Precision	Recall	Accuracy
83	Random Forest	Count Vectorizer (Bag of words method)	Noun – singular or mass + Noun – plural	0.996417	0.997505	0.997091
82	Random Forest	TF-IDF (Term Frequency & Inverse Document Frequency)	Noun – singular or mass + Noun – plural	0.995645	0.998284	0.997091
81	Random Forest	Term Frequency (TF – without IDF)	Noun – singular or mass + Noun – plural	0.995026	0.998284	0.996793
112	Adaptive Boosting	TF-IDF (Term Frequency & Inverse Document Frequency)	Noun – singular or mass + Noun – plural	0.995025	0.998129	0.996718
111	Adaptive Boosting	Term Frequency (TF – without IDF)	Noun – singular or mass + Noun – plural	0.994712	0.997505	0.996271
23	Logistic Regression	Count Vectorizer (Bag of words method)	Noun – singular or mass + Noun – plural	0.993767	0.994541	0.994406
113	Adaptive Boosting	Count Vectorizer (Bag of words method)	Noun – singular or mass + Noun – plural	0.993476	0.997505	0.995674
51	Passive Aggressive Classifier	Term Frequency (TF – without IDF)	Noun – singular or mass + Noun – plural	0.992389	0.996413	0.99463
66	Random Forest	Term Frequency (TF – without IDF)	Noun – plural	0.991628	0.997505	0.994779
67	Random Forest	TF-IDF (Term Frequency & Inverse Document Frequency)	Noun – plural	0.991318	0.997193	0.99448
78	Random Forest	Term Frequency (TF – without IDF)	Noun – singular or mass + Adjective	0.991162	0.979414	0.985977
68	Random Forest	Count Vectorizer (Bag of words method)	Noun – plural	0.991013	0.997505	0.99448
52	Passive Aggressive Classifier	TF-IDF (Term Frequency & Inverse Document Frequency)	Noun – singular or mass + Noun – plural	0.990983	0.994074	0.99284
53	Passive Aggressive Classifier	Count Vectorizer (Bag of words method)	Noun – singular or mass + Noun – plural	0.990963	0.99189	0.991795
80	Random Forest	Count Vectorizer (Bag of words method)	Noun – singular or mass + Adjective	0.990843	0.97879	0.98553
79	Random Forest	TF-IDF (Term Frequency & Inverse Document Frequency)	Noun – singular or mass + Adjective	0.99052	0.977698	0.984859
21	Logistic Regression	Term Frequency (TF – without IDF)	Noun – singular or mass + Noun – plural	0.99004	0.992202	0.991497
85	Random Forest	TF-IDF (Term Frequency & Inverse Document Frequency)	Noun – singular or mass + Adverb	0.989041	0.971148	0.981055
7	Logistic Regression	TF-IDF (Term Frequency & Inverse Document Frequency)	Noun – plural	0.988975	0.979258	0.984859

Random Forest gave the best results for the TF-IDF and Bag of Words method with Adaptive Boosting being the second using only the TF and Bag of Words method. Since both Random Forest and Adaboost are tree-based and combine weak learners, we will delve deeper with three models namely Random Forest with TF-IDF, Adaboost with TF and Logistic Regression with Bag of Words method.

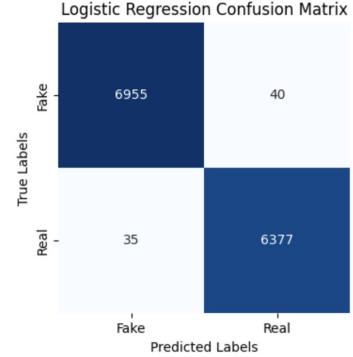
Random Forest with Bag of Words



AdaBoost with TF-IDF



Logistic Regression with Bag of Words



We tried a total of 120 (4 Models * 3 features * 10 filters) models out of which the Random Forest worked better than the AdaBoost and Logistic Regression model. Logistic Regression has slightly more misclassifications. Random Forest and AdaBoost were very close in prediction where each predicted only 9 and 12 instances respectively to be Fake when it was coming from the Real source. Logistic Regression missed 26 additional data points in predicting as Fake compared to Random Forest. Logistic Regression being a very simple model in comparison to Random Forest and Adaboost still performed a lot better and was able to capture the nuances and complex patterns in the dataset. It was off by an extra 15 records when predicting the data points as Real when it came from fake when compared. Logistic Regression is a linear model that makes predictions based on a linear combination of input features. Random Forest is a more complex model capable of capturing non-linear relationships. Random Forest has a higher tendency to overfit compared to Logistic Regression, especially when the dataset is noisy or contains irrelevant features.

Task 4 – Future Work

Please tell me what you plan to do in the future if you want to further enhance the performance of the machine learning models, e.g., enhancing learning models? Or investigating novel features? Before you propose your own idea, you may want to read the following papers:

1. [Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques](#)
2. [Supervised Learning for Fake News Detection](#)

Please design a simple experiment with GPT (or GPT API), and tell me if we can potentially use GPT to address this problem.

To further enhance the performance of machine learning models for fake news detection, here are some potential strategies for future plans:

- a. We can use "Feature Engineering" to explore and add more textual features such as readability scores, syntactic features, semantic similarity measures, specific publications that are well known, use of abusive context and exaggeration of events.
- b. We can make use of more sophisticated machine learning models such as deep learning architectures (e.g., LSTM, GRU, or transformers like BERT) to capture complex patterns and relationships in the text data.
- c. Since traditional machine learning models gave good results we can use "Ensemble Methods" to combine predictions from multiple models using ensemble techniques like bagging, boosting, or stacking to improve overall performance and robustness.
- d. We can include more types of news (apart from politics) using external data sources or use "Data Augmentation" to generate synthetic data or augment existing data to increase the diversity of the training set and use the actual data to increase the test set samples, which can help improve model generalization and performance.
- e. We can incorporate "domain-specific knowledge or external datasets" related to news articles, such as metadata, author credibility, publication source reputation, and historical context to diversify the training data extract more reliable features for our model and improve generalization capabilities.
- f. Since "Hyperparameter Tuning" we can even try this method as well as K-Fold Cross Validation methods to gauge the reliability of our model. To perform hyperparameter tuning using techniques like grid search, random search or Bayesian optimization to find the optimal set of parameters for the models.

Regarding the use of ChatGPT to address the fake news detection problem:

ChatGPT can be utilized in the following ways:

- a. We can use ChatGPT as "Text Classification Assistance" to provide context-based suggestions or explanations for the classification decisions made by the models, helping users understand the rationale behind the predictions.
- b. We can use ChatGPT as "Adversarial Testing Assistance" to generate adversarial examples or fake news samples with the intention of testing the robustness of the classification models and identifying potential vulnerabilities.
- c. We can use ChatGPT as "Fact-Checking and Verification Assistance" to assist in fact-checking and verifying the authenticity of news articles by providing relevant information or cross-referencing against credible sources. Given its extensive training on a diverse dataset, ChatGPT can effectively aid in verifying whether descriptions discuss an exaggeration of events through classification, thus serving as a valuable resource for fact-checking and verification assistance in news articles.

Simple Experiment with ChatGPT

Initial Prompt

Assume CHATGPT is trained to classify into fake or real based on the description. what key factors will CHATGPT look into before making its final call and judgment? Describe your answer in brief and list down all the key points.



ChatGPT

If ChatGPT is trained to classify text into fake or real news, it would consider several key factors before making its final judgment. These factors include:

1. **Language Patterns:** ChatGPT would analyze the language patterns and structures used in the text, including grammar, syntax, and semantics.
2. **Contextual Understanding:** It would consider the context of the text, including the topic, tone, and coherence of the content.
3. **Factual Accuracy:** ChatGPT would assess the factual accuracy of the information presented in the text by cross-referencing against credible sources or known facts.
4. **Source Reliability:** It would evaluate the reliability and credibility of the source from which the text originated, including the publication's reputation and history of accurate reporting.
5. **Consistency:** ChatGPT would check for consistency within the text and compare it with known factual information or similar news articles.
6. **Bias and Sensationalism:** It would identify any biases or sensationalized language that could indicate an agenda or manipulation of information.
7. **Corroboration:** ChatGPT might look for corroborating evidence or multiple sources supporting the claims made in the text.
8. **Contextual Nuances:** It would consider contextual nuances such as cultural references, political affiliations, and historical background to understand the underlying message.
9. **Deceptive Techniques:** ChatGPT would be trained to recognize common deceptive techniques used in fake news, such as misinformation, propaganda, or misleading headlines.
10. **Critical Thinking:** It would apply principle ↓ critical thinking to assess the validity of arguments, logical fallacies, and evidence presented in the text.

Testing ChatGPT by giving FAKE news for classification.

this is a piece of text classify it as either fake or real and describe your steps in the classification.
news text starts from here

:::::

donald trump sends embarrassing new year ' eve message ; disturbing donald trump wish american happy new year leave . instead , give shout enemy , hater dishonest fake news medium . former reality show star one job . country rapidly grows stronger smarter , want wish friend , supporter , enemy , hater , even dishonest fake news medium , happy healthy new year , president angry pant tweeted . 2018 great year america ! country rapidly grows stronger smarter , want wish friend , supporter , enemy , hater , even dishonest fake news medium , happy healthy new year . 2018 great year america ! donald j. trump (@ realdonaldtrump) december 31 , 2017trump tweet went welll expect.what kind president sends new year greeting like despicable , petty , infantile gibberish ? trump ! lack decency even allow rise gutter long enough wish american citizen happy new year ! bishop talbert swan (@ talbertswan) december 31 , 2017no one like calvin (@ calvinstowell) december 31 , 2017your impeachment would make 2018 great year america , also accept regaining control congress . miranda yaver (@ mirandayaver) december 31 , 2017do hear talk ? include many people hate wonder ? hate ? alan sandoval (@ alansandoval13) december 31 , 2017who us word hater new year wish ?? marlene (@ marlene399) december 31 , 2017you say happy new year ? koren pollitt (@ korencarpenter) december 31 , 2017here trump new year eve tweet 2016.happy new year , including many enemy fought lost badly know . love ! donald j. trump (@ realdonaldtrump) december 31 , 2016this nothing new trump . years.trump directed message enemy hater new year , easter , thanksgiving , anniversary 9/11 . pic.twitter.com/4fpae2kypa daniel dale (@ ddale8) december 31 , 2017trump holiday tweet clearly presidential.how long work hallmark becoming president ? steven goodine (@ sgoodine) december 31 , 2017he always like .. . difference last year , filter breaking . roy schulze (@ thbthttt) december 31 , 2017who , apart teenager us term hater ? wendy (@ wendywhistles) december 31 , 2017he fucking 5 year old know (@ rainyday80) december 31 , 2017so , people voted hole thinking would change got power , wrong ! 70-year-old men change year older.photo andrew burton/getty image .

:::::

news text ends here



ChatGPT



To classify the given news text as either fake or real, I would follow these steps:

To classify the given news text as either fake or real, I would follow these steps:

- .. **Read and Understand the Text:** Carefully read and understand the content of the news text provided.
2. **Identify Key Indicators:** Look for key indicators that might suggest whether the text is fake or real. This could include the tone of the language, the credibility of the sources mentioned, and the presence of factual inaccuracies or biased language.
3. **Analyze Language Patterns:** Analyze the language patterns used in the text, including grammar, syntax, and vocabulary. Pay attention to any inconsistencies or unusual language choices that could indicate potential manipulation or deception.
4. **Check Source and Context:** Investigate the source of the news text and consider the broader context in which it was published. Evaluate the credibility and reliability of the source, and assess whether the content aligns with known facts or events.
5. **Verify Claims:** Fact-check any claims or assertions made in the text by cross-referencing with reputable sources or conducting additional research.
6. **Consider External Factors:** Take into account any external factors or biases that might influence the content of the news text, such as political affiliations, agendas, or conflicts of interest.
7. **Make Judgment:** Based on the analysis of the language, sources, context, and verifiability of claims, make a judgment about whether the news text is likely to be fake or real.

Based on the provided news text, it appears to contain derogatory language and personal attacks, which could indicate a biased or sensationalized tone. Additionally, the lack of credible sources and the presence of inflammatory language suggest that the text may be fake rather than real news. However, further analysis and fact-checking would be necessary to make a definitive classification.

The prediction was correct as it was a fake text from the corpus.

Testing ChatGPT by giving REAL news for classification.

this is a piece of text classify it as either fake or real and describe your steps in the classification.
news text starts from here

:::::

u.s. budget fight loom , republican flip fiscal script washington (reuters) - head conservative republican faction u.s. congress , voted month huge expansion national debt pay tax cut , called “ fiscal conservative ” sunday urged budget restraint 2018. keeping sharp pivot way among republican , u.s. representative mark meadow , speaking cbs ‘ “ face nation , ” drew hard line federal spending , lawmaker bracing battle january . return holiday wednesday , lawmaker begin trying pas federal budget fight likely linked issue , immigration policy , even november congressional election campaign approach republican seek keep control congress . president donald trump republican want big budget increase military spending , democrat also want proportional increase non-defense “ discretionary ” spending program support education , scientific research , infrastructure , public health environmental protection . “ (trump) administration already willing say : ‘ ‘ going increase non-defense discretionary spending ... 7 percent , ’ ” meadow , chairman small influential house freedom caucus , said program . “

:::::

news text ends here



ChatGPT

To classify the provided news text as either fake or real, I'll follow these steps:

1. **Read and Understand the Text:** Carefully read and understand the content of the news text provided.
2. **Identify Key Indicators:** Look for key indicators that might suggest whether the text is fake or real. This could include the language used, the credibility of the sources mentioned, and the presence of factual information.



ChatGPT

To classify the provided news text as either fake or real, I'll follow these steps:

1. **Read and Understand the Text:** Carefully read and understand the content of the news text provided.
2. **Identify Key Indicators:** Look for key indicators that might suggest whether the text is fake or real. This could include the language used, the credibility of the sources mentioned, and the presence of factual information.
3. **Analyze Language Patterns:** Analyze the language patterns used in the text, including grammar, syntax, and vocabulary. Pay attention to any inconsistencies or unusual language choices that could indicate potential manipulation or deception.
4. **Check Source and Context:** Investigate the source of the news text and consider the broader context in which it was published. Evaluate the credibility and reliability of the source, and assess whether the content aligns with known facts or events.
5. **Verify Claims:** Fact-check any claims or assertions made in the text by cross-referencing with reputable sources or conducting additional research.
6. **Consider External Factors:** Take into account any external factors or biases that might influence the content of the news text, such as political affiliations, agendas, or conflicts of interest.
7. **Make Judgment:** Based on the analysis of the language, sources, context, and verifiability of claims, make a judgment about whether the news text is likely to be fake or real.

Based on the provided news text, it appears to be reporting on political events related to the U.S. budget, including statements from conservative Republican representatives and discussions about government spending. Without further context or verification, it's difficult to determine definitively whether the news text is fake or real. However, the presence of specific details and quotes from public figures suggests that the text may be reporting on actual events, rather than being fabricated or misleading. Further analysis and fact-checking would be necessary to make a definitive classification.

The prediction was correct as it was a real text from the corpus.