PROJECT ON FAKE NEWS DETECTION

Submitted by:
SHALINI ROY

FlipRobo SME:
KHUSHBOO GARG

ACKNOWLEDGMENT

I want to thank the Flip Robo Technologies team for giving me the chance to work with this dataset during my internship. It enabled me to develop my analytical abilities.The entire DataTrained team deserves a great thank you.

Reference used in this project:

- GitHub Notes & Repository.
- Various Kaggle and Github projects.
- Analytics Vidya's different papers on Data Science.
- SCIKIT Learn Library Documentation.
- Predicting from www.cardekho.com

Abstract

The spread of fake news on social media and other platforms is a serious concern because it has the potential to have a negative impact on society and the country. On finding it, there has already been a lot of research. In order to develop a model of a product with supervised machine learning algorithm, which can classify fake news as true or false by using tools like Python Scikit-Learn, NLP for textual analysis, this paper analyses the research on fake news detection and explores the best traditional machine learning models.

We suggest using the Python scikit-learn library to perform tokenization and feature extraction of text data because it contains practical tools like the Count Vectorizer and Tiff Vectorizer. This process will result in feature extraction and vectorization. Then, based on the results of the confusion matrix, we will use feature selection techniques to experiment and select the best-fit features to achieve the highest precision.

Introduction

What is fake news?

The definition of fake news is to include information that misdirects readers. These days, fake news spreads like wildfire, and people spread it without checking the facts. This is frequently accomplished with political agendas in order to advance or impose particular ideas. To make money from online advertising, media outlets must be able to draw viewers to their websites. Therefore, it's important to spot fake news.

False information can be found in fake news and could be verified. This perpetuates a lie about a particular statistic in a nation or inflates the cost of a particular service for a nation, which may cause unrest in some nations, like the Arab Spring. There are groups working to address issues like verifying authors' accountability, such as the House of Commons and the Crosscheck project. However, because they rely on manual detection by humans, which is impossible to control or implement manually in a world where millions of articles are either removed or published every minute. The creation of a system to provide a reliable automated index scoring, or rating, for the credibility of various publishers, and the context of the news, could be a solution.

Natural Language Processing

Natural Language Processing is primarily used to take into account one or more system or algorithm specializations. Speech understanding and speech generation can be combined using an algorithmic system's Natural Language Processing (NLP) rating. It could also be used to track actions in different languages. Emotion Analyzer and Detection, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Chunking, and Semantic Role Labeling made NLP a good Subject of the search and suggested a new ideal system for extraction actions from languages of English, Italian, and Dutch speeches through the use of various pipelines of various languages.

Sentiment analysis collects feelings about a specific topic. Extraction of a specific term for a subject, extraction of the sentiment, and coupling with connection analysis make up sentiment analysis. Sentiment analysis employs bilingualism A few sources for research Dictionary of terms and library of sentiment models. for positive and negative terms and makes an effort to categorise them on a scale of -5 to 5. To create parts of language taggers for languages like Sanskrit, Hindi, and Arabic, researchers are looking into parts of speech taggers for languages like European languages.
It may be effective Mark and classify words as adjectives, verbs, names, and so on. The majority of part-of-speech approaches work well in European languages but not in Asian or Arabic.

Inserting a column called 'Class' in Fake and true datasets:

```
In [5]: df_fake["class"] = 0
        df_true["class"] = 1

In [6]: df_fake.shape, df_true.shape

Out[6]: ((23481, 5), (21417, 5))
```

Checking Null:

```
In [11]: null_val= df_true.isna().sum().any()
         null_val

Out[11]: False

In [12]: null_val= df_fake.isna().sum().any()
         null_val

Out[12]: False
```

There is no null values

- No, Null value is present.
- Removing the last 10 rows from both datasets, for manual testing.
- Merging the manual testing data frame in a single dataset and saving it in a CSV file.
- "title", "subject" and "date" columns is not required for detecting fake news, so I am going to drop the columns.

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods):

In this problem regression-based machine learning algorithm like linear regression can be used. For that first data encoding and data scaling using standard scalar is done. For building an appropriate ML model before implementing classification algorithms, data is split in training & test data using train_test_split.

Testing of Identified Approaches (Algorithms)

Total of 4 algorithms used for the training and testing are:
1. Linear Regression
2. Decision Tree Classification
3. Gradient Boosting Classifier
4. Random Forest Classifier

Linear Regression:

```
In [42]: LR.score(xv_test, y_test)
Out[42]: 0.9878787878787879

In [43]: print(classification_report(y_test, pred_lr))
```

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5856
1	0.99	0.99	0.99	5364
accuracy			0.99	11220
macro avg	0.99	0.99	0.99	11220
weighted avg	0.99	0.99	0.99	11220

Accuracy coming as 99

Decision Tree Classification:

```
In [47]: DT.score(xv_test, y_test)
```

```
Out[47]: 0.9958110516934047
```

```
In [48]: print(classification_report(y_test, pred_dt))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5856
           1       1.00      1.00      1.00      5364

    accuracy                           1.00     11220
   macro avg       1.00      1.00      1.00     11220
weighted avg       1.00      1.00      1.00     11220
```

Decision Tree Classification Accuracy coming almost 100

Gradient Boosting Classifier:

```
In [52]: GBC.score(xv_test, y_test)
```

```
Out[52]: 0.9958110516934047
```

```
In [53]: print(classification_report(y_test, pred_gbc))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      5856
           1       0.99      1.00      1.00      5364

    accuracy                           1.00     11220
   macro avg       1.00      1.00      1.00     11220
weighted avg       1.00      1.00      1.00     11220
```

Gradient Boosting Classifier Accuracy as 99.5

Random Forest Classifier:

```
In [57]: RFC.score(xv_test, y_test)
```

```
Out[57]: 0.9909982174688057
```

```
In [58]: print(classification_report(y_test, pred_rfc))
```

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5856
           1       0.99      0.99      0.99      5364

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

Random Forest Classifier Accuracy as 99

Model Testing with Manual Entry

We have done two manual testing and the result are below:

```
news = str(input())
manual_testing(news)
```

```
WEST PALM BEACH, Fla./WASHINGTON (Reuters) - The White House said on Friday it was set to kick off talks next week with Republi
can 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 fo
r midterm congressional elections in November. President Donald Trump's Republican Party is eager to maintain control of Congre
ss 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 an
d House Speaker Paul Ryan - both Republicans - and their Democratic counterparts, Senator Chuck Schumer and Representative Nanc
y Pelosi, the White House said. That will be followed up with a weekend of strategy sessions for Trump, McConnell and Ryan on J
an. 6 and 7 at the Camp David presidential retreat in Maryland, according to the White House. The Senate returns to work on Ja
n. 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 dow
n. Also on the agenda for lawmakers is disaster aid for people hit by hurricanes in Puerto Rico, Texas and Florida, and by wild
fires 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 ente
red the country illegally as children, known as "Dreamers." In September, Trump ended Democratic former President Barack Obam
a'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 larg
e companies have pushed for DACA protections to continue. Trump and other Republicans have said that will not happen without Co
ngress approving broader immigration policy changes and tougher border security. Democrats oppose funding for a wall promised b
y 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 Im
migration etc," Trump said in a Twitter post on Friday. Trump wants to overhaul immigration rules for extended families and oth
ers 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 DRE
AMers, 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 ear
ly April if Congress does not raise the debt ceiling before then, according to the nonpartisan Congressional Budget Office. Tru
mp, who won his first major legislative victory with the passage of a major tax overhaul this month, has also promised a major
infrastructure plan.


LR Prediction: Not A Fake News
DT Prediction: Not A Fake News
GBC Prediction: Not A Fake News
RFC Prediction: Not A Fake News
```

➢ **This news has been identified as "Not A Fake News"**

```
In [61]: news = str(input())
         manual_testing(news)
```

```
On Christmas day, Donald Trump announced that he would  be back to work  the following day, but he is golfing for the fourth da
y in a row. The former reality show star blasted former President Barack Obama for playing golf and now Trump is on track to ou
tpace the number of golf games his predecessor played.Updated my tracker of Trump s appearances at Trump properties.71 rounds o
f golf including today s. At this pace, he ll pass Obama s first-term total by July 24 next year. https://t.co/Fg7VacxRtJ pic.t
witter.com/5gEMcjQTbH  Philip Bump (@pbump) December 29, 2017 That makes what a Washington Post reporter discovered on Trump s
website really weird, but everything about this administration is bizarre AF. The coding contained a reference to Obama and gol
f:  Unlike Obama, we are working to fix the problem  and not on the golf course.  However, the coding wasn t done correctly.Th
e website of Donald Trump, who has spent several days in a row at the golf course, is coded to serve up the following message i
n the event of an internal server error: https://t.co/zrWpyMXRcz pic.twitter.com/wiQSQNNzw0  Christopher Ingraham (@_cingraham)
December 28, 2017That snippet of code appears to be on all https://t.co/dkhw0AlHB4 pages, which the footer says is paid for by
the RNC? pic.twitter.com/oaZDT126B3  Christopher Ingraham (@_cingraham) December 28, 2017It s also all over https://t.co/ayBlGm
k65Z. As others have noted in this thread, this is weird code and it s not clear it would ever actually display, but who knows.
Christopher Ingraham (@_cingraham) December 28, 2017After the coding was called out, the reference to Obama was deleted.UPDATE:
The golf error message has been removed from the Trump and GOP websites. They also fixed the javascript  =  vs  ==  problem. St
ill not clear when these messages would actually display, since the actual 404 (and presumably 500) page displays a different m
essage pic.twitter.com/Z7dmyQ5smy  Christopher Ingraham (@_cingraham) December 29, 2017That suggests someone at either RNC or t
he Trump admin is sensitive enough to Trump s golf problem to make this issue go away quickly once people noticed. You have no
idea how much I d love to see the email exchange that led us here.  Christopher Ingraham (@_cingraham) December 29, 2017 The co
de was f-cked up.The best part about this is that they are using the  =  (assignment) operator which means that bit of code wil
l never get run. If you look a few lines up  errorCode  will always be  404          (@tw1trsux) December 28, 2017trump s coder
s can t code. Nobody is surprised.  Tim Peterson (@timrpeterson) December 28, 2017Donald Trump is obsessed with Obama that his
name was even in the coding of his website while he played golf again.Photo by Joe Raedle/Getty Images.


LR Prediction: Fake News
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News
```

➢ **This news has been identified as "Fake News"**

Saving the Model

Saving the model with the pickle method for future reference.

In [65]:
```
# Saving the best model.
import pickle
pickle.dump(GBC.predict,'Fake_news_Project.pkl')
```

Out[65]:
```
['Fake_news_Project.pkl']
```

Thank you!!