FAKE NEWS DETECTION

Report

NOVEMBER 17, 2021

LOVELY PROFESSIONAL UNIVERSITY
Jalandhar,Punjab

Final Project Report

Fake News Detection

Deepshikha Singh

12001693

A report submitted in part fulfilment of the degree of

B. Tech in Computer Science

Faculty Name-Ankita Wadhawan

School of Computer Science and Engineering

Lovely Professional University

November 17, 2021

Contents

Abstract

The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in the human history before. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality.

Several studies have primarily focused on detection and classification of fake news on social media platforms such as Facebook and Twitter. At conceptual level, fake news has been classified into different types; the knowledge is then expanded to generalize machine learning (ML) models for multiple domains. This paper makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python NumPy, pandas for analysis. This process will result in feature extraction and vectorization; we propose using Python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer. Then, we will perform feature selection methods, to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results.

Acknowledgment

Presentation inspiration and motivation have always played a key role in the success of any venture.

I express my sincere thanks to Ankita Wadhawan Ma’am, faculty, Lovely Professional University. Who gave me the golden opportunity to do this wonderful project of Python.

Who also helped me in completing my project. I came to know about so many new things I am thankful to her.

DECLARATION

I the undersigned solemnly declare that the project report is based on my own work carried out during our study under the supervision of Miss Ankita Wadhawan.

I assert the statements made and conclusions drawn are an outcome of my project work. I further certify that

I. The work contained in the report is original and has been done by me under the general supervision of my supervisor.
II. I have followed the guidelines provided by the university in writing the report.
III. Whenever I have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and giving their details in the references.

Deepshikha Singh

12201693

INTRODUCTION

Fake news refers to information content that is false, misleading or whose source cannot be verified. This content may be generated to intentionally damage reputations, deceive, or to gain attention.

Various types of fake news include:

- Clickbait. Often eye-catching content to capture readers at the expense of being factual.
- Satire/parody. This type of content is fun and humorous thus considered to be entertaining, yet some readers may interpret the content as fact.
- Propaganda. This is content meant to mislead and influence the reader.
- Biased/partisan/hyper-partisan. Oftentimes this is biased political content claiming to be impartial.
- Unreliable news. Journalists may publish news whose sources are unverified, or without carrying out any form of fact checking themselves.

How Fake News Works

Social media platforms are incredibly influential. The estimated daily number of tweets is about 500 million. These platforms are ubiquitous. They are the go-to environment to share thoughts, feelings, opinions, and intentions. This provides ideal conditions to distribute news with minimal guidelines and restrictions.

In today's world, it is normal to receive news from online sources like social media. News is often subjective to readers. We often choose to ingest content that appeals to the different emotions we have. So, considering this, the information that gets the most reach may not be real or accurate news.

Additionally, real news may be twisted in transmission. A reader may end up with different versions of the same news. This may lead to information overload.

What's being done to combat fake news

Companies like Facebook, Twitter, TikTok, Google, Pinterest, Tencent, YouTube, and others are working with WHO to mitigate the spread of rumours. Their efforts are geared at filtering out content that is a danger to public health. There are ways to contribute to this fight. But first, we need to understand the types of fake news detection being used. We will look at it from the perspective of being either manual or automatic.

Manual Fake News Detection

Manual fake news detection often involves all the techniques and procedures a person can use to verify the news. It could involve visiting fact checking sites. It could be crowdsourcing real news to compare with unverified news. But, the amount of data generated online daily is overwhelming. Also noting how fast information spreads online, manual fact checking quickly becomes ineffective. Manual fact checking struggles to scale with the volume of data generated. Therefore, highlighting the reason behind the creation of automated fake news detection.

Automated Fake News Detection

Automated detection systems provide value in terms of automation and scalability. There are various techniques and approaches implemented in fake news detection research. And it is worth noting that these approaches often overlap depending on perspective.

The two approaches to fake news detection are:

- Machine Learning approach
- Deep Learning approach

Machine Learning Approach

Machine learning refers to giving computers the ability to learn without explicitly being programmed. A machine learning approach uses machine learning algorithms to detect misinformation. Examples of these algorithms include:

Decision Tree: a supervised learning algorithm that has a tree-like flow. It helps in decision making. A useful algorithm for both classification and regression tasks.

Support Vector Machine: a supervised learning algorithm. It examines data for classification and regression analysis. It classifies data into two categories.

Logistic Regression: contrary to the name, it is a classification algorithm used to estimate discrete values.

LIBRARIES

NumPy:

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices.

In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.

The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy. Arrays are very frequently used in data science, where speed and resources are very important.

Pandas:

Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analysing data much easier. Pandas is fast and it has high-performance & productivity for users.

Pandas makes it simple to do many of the time consuming, repetitive tasks associated with working with data, including:

- Data cleansing
- Data fill
- Data normalization
- Merges and joins
- Data visualization
- Statistical analysis
- Data inspection
- Loading and saving data
- And much more

Scikit Learn (Sklearn):

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression,

clustering and dimensionality reduction via a consistence interface in Python.

Data Set:

The dataset we'll use for this python project- we'll call it news.csv. This dataset has a shape of 3988*3. The first column identifies the news, the second and third are the title and text, and the fourth column has labels denoting whether the news is REAL or FAKE.

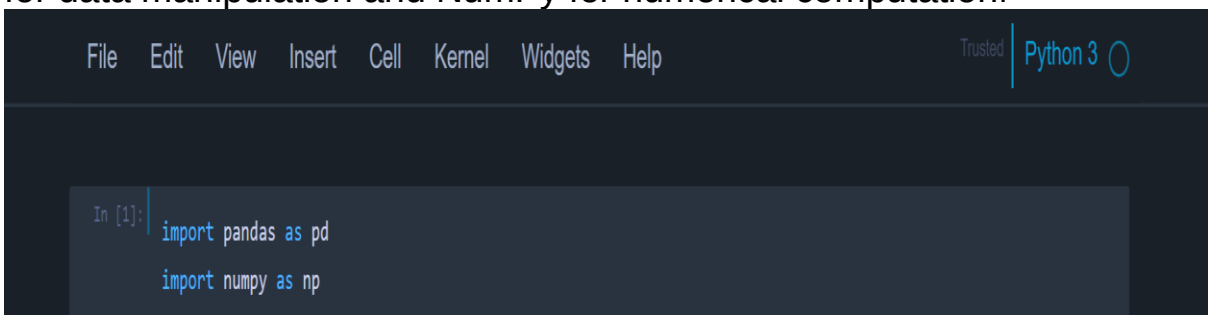
DECISION TREE:

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. The intuition behind **Decision Trees** is that you use the dataset features to create *yes/no* questions and continually split the dataset until you isolate all data points belonging to each class

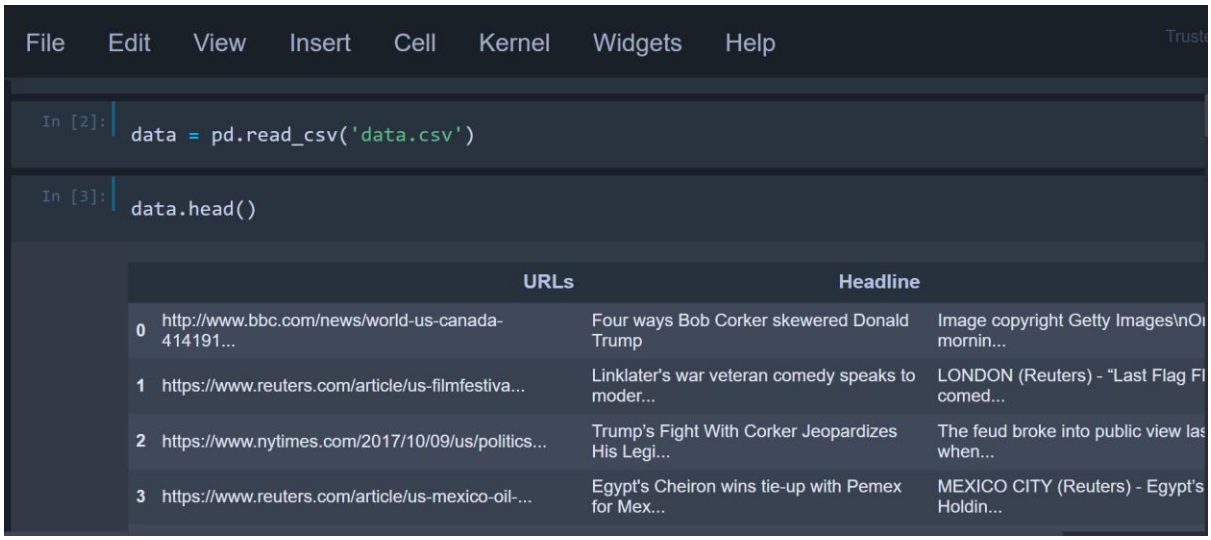
CODE EXPLANATION

First, I imported two libraries called pandas and NumPy, Using pandas for data manipulation and NumPy for numerical computation.



After that, **used read_csv, which** is an important pandas function to read csv files and do operations on it.

For the glance of data set, data.head is used



As we do not require URLs column, so using drop command, we dropped this column and with this also dropped other empty values.

File Edit View Insert Cell Kernel Widgets Help

In [4]:
```
data = data.drop(['URLs'],axis=1)
data = data.dropna()
```

In [5]:
```
data.head()
```

	Headline	Body	Label
0	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	1
1	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - “Last Flag Flying”, a comed...	1
2	Trump’s Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	1
3	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt’s Cheiron Holdin...	1
4	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1

After that we are extracting dependent and independent variables and storing them into x and y.

In [6]:
```
#dependent and indenpendent variables
x = data.iloc[:,:-1].values #independent
y = data.iloc[:,-1].values#dependent
```

In [7]:
```
#First record of independent variables
x[0]
```

```
array(['Four ways Bob Corker skewered Donald Trump',
       'Image copyright Getty Images\nOn Sunday morning, Donald Trump went off on a Twitter tirade against a member of
s, in itself, isn\'t exactly huge news. It\'s far from the first time the president has turned his rhetorical cannons o
his time, however, his attacks were particularly biting and personal. He essentially called Tennessee Senator Bob Corke
powerful Senate Foreign Relations Committee, a coward for not running for re-election.\nHe said Mr Corker "begged" for
```

Now, converting all the text data into numerical form using Count Vectorizer. Then converting I into dense matrix.

After converting, stacking headlines on the top of body .

File Edit View Insert Cell Kernel Widgets Help

```
#Converting text data into numerical
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=5000)
mat_body = cv.fit_transform(x[:,1]).todense() #converting into dense matrix
```

In [10]:

```
mat_body
```

```
matrix([[0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        ...,
        [0, 2, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

File Edit View Insert Cell Kernel Widgets Help

```
cv_head = CountVectorizer(max_features=5000)
mat_head = cv_head.fit_transform(x[:,0]).todense()
```

In [12]:

```
mat_head
```

```
matrix([[0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        ...,
        [1, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [13]:

```
#stacking headlines on top of body
x_mat = np.hstack((mat_body, mat_head))
```

Now dividing data into test and train, importing test_train_split from Sklearn library. We are taking test_size=0.2 which means 20% of record would be in test set and other 80% training set would be in train set.

Now we have test and training set ready, now we finally go ahead and build **Decision Tree Classifier**

```
#stacking headlines on top of body
x_mat = np.hstack((mat_body, mat_head))
```

In [14]:

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_mat, y, test_size=0.2, random_state=0)
```

In [15]:

```
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier(criterion='entropy')
dtc.fit(x_train, y_train)
y_pred_dtr=dtc.predict(x_test)
```

After that build a confusion matrix, left diagonal will predict all the correctly predicted news.

In [16]:
```
#making a confusion matrix to check the accuracy
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test,y_pred_dtr )
```

```
array([[426,   8],
       [ 18, 346]], dtype=int64)
```

In [17]:
```
(422+12)/(422+12+17+347) #correctly predicted result divided by all of the results
```

```
0.543859649122807
```

CONCLUSION

Fake news research has never been more important than it is now. Especially during a time when the world is fighting a pandemic. There are so many more approaches and criteria for fake news detection. Datasets also impact the accuracy of fake news detection tasks.

Their quality and quantity are impactful. It is also worth noting that, as much as our focus is on automated approaches, the human element is key to this fight. A combination of human and automated approaches gives rise to a hybrid approach.

Using decision tree classifier in this project, accuracy rate of correctly predicted news is 54%.

GITHUB LINK

https://github.com/DeepshikhaSingh18/Fake-News-Detection