Fake News Detection On Social-Media

A PROJECT REPORT

for

DATA MINING TECHNIQUES (ITE2006)

in

B.Tech. (IT)

by

MEGHA SAXENA (20BIT0366)

Fourth Semester, 2022

Under the Guidance of

Prof. VALARMATHI B

Associate Professor (Senior), SITE

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

May-June, 2022

DECLARATION BY THE CANDIDATE

We here by declare that the project report entitled **"Fake News Detection On Social-Media"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide project work carried out by us under the guidance of **Prof. VALARMATHI B.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore

Signature

Date : 28-04-2022

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled **"Fake News Detection On Social-Media"** submitted by **MEGHA SAXENA (20BIT0366)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide work carried out by them under my guidance.

Prof. VALARMATHI B

GUIDE
Asso. Professor(senior), SITE

Fake News Detection On Social-Media

BY: MEGHA SAXENA (20BIT0366)

Objective

The objective of this project is to identify fake news using surveys and existing algorithms from data mining. In this survey, I will present a comprehensive review of detecting fake news on social media using data mining techniques. I will also present a comprehensive review of detecting fake news on social media, including fake news characterizations. The project will also see related research areas, open problems, and future research directions for fake news detection on social media.

Abstract

Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and fast dissemination of data lead humans to seek out and consume information from social media. On the other hand, it allows the extensive unfold of “faux information”, i.e., low-quality news with intentionally false information. The extensive spread of faux information has the potential for extremely poor effects on people and society. Fake news detection on social media presents specific traits and demanding situations that make current detection algorithms from conventional information media useless or now no longer applicable. In this project, I will identify fake news using data mining techniques.

KEYWORD- Fake News, social media, data mining, logistic regression

INTRODUCTION

Fake news detection on social media presents specific traits and demanding situations that make current detection algorithms from conventional information media useless or now no longer applicable. First, faux information is deliberately written to lie to readers to trust fake information, which makes it hard and nontrivial to discover primarily based totally on news content; therefore, we want to include auxiliary information, which includes user social engagements on social media, to assist make a determination. Second, exploiting this auxiliary information is difficult in and of itself as users’ social engagements with faux information produce data that is big, incomplete, unstructured, and noisy.

Literature survey

S. No	Title of paper (YEAR)	Algorithm Used	Data set being used	Performance measures	Scope for future work

1	Fake News Detection on Social Media: A Data Mining Perspective (2017)	Given the social news engagements among n users for news article a, the task of fake news detection is to predict whether the news article is fake news piece or not.	Buzzfeed	Accuracy-70-75%	A promising direction is to create a comprehensive and large-scale fake news benchmark dataset, which can be used by researchers to facilitate further research in this area
2	Fake news detection in social media (2018)	a combination of Naïve Bayes classifier, Support Vector Machines, and semantic analysis	Data is taken from (Shu, Sliva, Wang, Tang, & Liu, 2017)	Accuracy-80%	This research may be used to help other researchers discover which combination of methods should be used in order to accurately detect fake news in social media.
3	Detection of fake news using deep learning CNN–RNN based methods	KNN Classifier	Liar Dataset	Accuracy-95%	Comparing the accuracies would be beneficial in deciding whether or not the dataset is representative of how difficult the task of separating fake from real news is.
4	Analysis of Classifiers for Fake News Detection (2019)	The performance of a classifier may vary based on the size and quality of the text data (or corpus) and also the features of the text vectors. Common noisy wordscalled ‘stopwords’ are less important words when it comes to text feature extraction, they don’t contribute towards the actual meaning of a sentence and they only contribute towards feature dimensionality and may be discarded for better performance.	Self Surveys	Accuracy-75%	For future improvements, concepts like POS tagging, word2vec and topic modelling can be utilized. These will give the model a lot more depth in terms of feature extraction and fine-tuned classification.
5	An Efficient Supervised Method for Fake News	NaiveBayes Algorithm	Collected from various sources	Accuracy-94.6%	Collection the classifiers to attain higher performance

	Detection using Machine and Deep Learning Classifiers (2020)				victimisation ADA Boost methodology
6	AUTOMATIC FAKE NEWS DETECTION (2020)	To extract temporal representations of articles we use a Recurrent Neural Network (RNN). Temporal engagements are stored as vectors and are fed into the RNN which produces an output a representation vector vj .	Google Scholar	Accuracy-89%	One particularly interesting direction would be to build models that incorporate concepts from reinforcement learning and crowd sourcing. Including humans in the learning process could lead to more accurate and, in particular, more timely predictions
7	Fake News Detection Using Machine Learning Approaches (2021)	• Random Forests. • Naive Bayes. • K-Nearest Neighbors (KNN). • Decision Tree. • SVM	Scholar Space	Accuracy-More than 75%	fake news detection approaches that is based on text analysis in the paper utilizes models based on speech characteristics and predictive models that do not fit with the current models
8	Fake News Detection with Naïve Bayes Classifier (2021)	NaiveBayes Algorithm.	Kaggle	Accuracy-95%	Through further research, this kind of accuracy can be achieved using more classifiers
9	Fake News Detection using Machine Learning (2021)	Multinomial Naive Bayes algorithm is a probabilistic learning methodology that's principally utilized in Natural Language Processing (NLP). The algorithmic program relies on the Bayes theorem and predicts the tag of a text like a bit of email or newspaper article. It calculates the probability of each tag	Kaggle	Accuracy-87%	This model can be more applicable in the future for other regional languages (Like Hindi, Marathi, Bengali, etc.) and especially using a native country dataset by either collecting data using a Twitter/Google/Reddit API or by web scraping from a Social Media /News Website

		for a given sample and then provides the tag with the highest probability as output			
10	Sentiment Analysis for Fake News Detection (2021)	NLP Algorithm	Ad-hoc	Accuracy-85%	Direct comparison between systems and approaches is so far difficult due to the wide range of data sets used, many of them ad hoc

TABLE 1: Literature Survey

Existing system

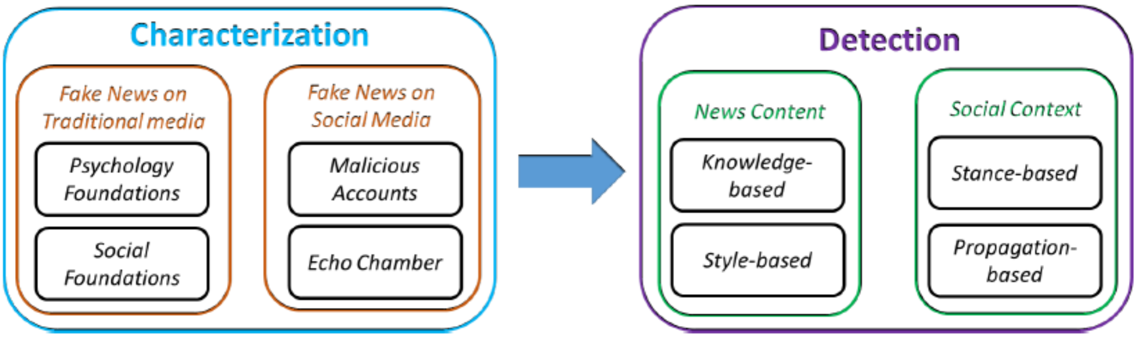

FIGURE 1: Fake news on social media: from characterization to detection

CHARACTERIZATION

Fake news on traditional media

Psychological Foundations of Fake News

Humans are naturally not very good at differentiating between real and fake news. There are several psychological and cognitive theories that can explain this phenomenon and the influential power of fake news. Traditional fake news mainly targets consumers by exploiting their individual vulnerabilities.

Social Foundations of the Fake News Ecosystem

Considering the entire news consumption ecosystem, we can also describe some of the social dynamics that contribute to the proliferation of fake news. Prospect theory describes decision making as a process by which people make choices based on the relative gains and losses as compared to their current state

Fake news on social media

Malicious Accounts on Social Media for Propaganda

While many users on social media are legitimate, social media users may also be malicious, and in some cases are not even real humans. The low cost of creating social media accounts also encourages malicious user accounts, such as social bots, cyborg users, and trolls.

Echo Chamber

Social media provides a new paradigm of information creation and consumption for users. The information seeking and consumption process are changing from a mediated form (e.g., by journalists) to a more disinter-mediated way.

DETECTION

To evaluate the performance of algorithms for fake news detection problem, various evaluation metrics have been used.

In this subsection, we review the most widely used metrics for fake news detection. Most existing approaches consider the fake news problem as a classification problem that predicts whether a news article is fake or not:

True Positive (TP): when predicted fake news pieces are actually annotated as fake news;
True Negative (TN): when predicted true news pieces are actually annotated as true news;
False Negative (FN): when predicted true news pieces are actually annotated as fake news;
False Positive (FP): when predicted fake news pieces are actually annotated as true news.
By formulating this as a classification problem, we can define following metrics,

$$\text{Precision} = |TP|/(|TP|+|FP|)$$
$$\text{Recall} = |TP|/(|TP|+|FN|)$$
$$F1 = 2\{(\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall})\}$$
$$\text{Accuracy} = (|TP|+|TN|)/(|TP|+|TN|+|FN|+|FP|)$$

Features / Dataset	News Content		Social Context		
	Linguistic	Visual	User	Post	Network
BuzzFeedNews	✓				
LIAR	✓				
BS Detector	✓				
CREDBANK	✓		✓	✓	✓

FIGURE 2: Comparison of Fake news detection Dataset

These metrics are commonly used in the machine learning community and enable us to evaluate the performance of a classifier from different perspectives. Specifically, accuracy measures the similarity between predicted fake news and real fake news. Precision measures the fraction of all detected fake news that are annotated as fake news, addressing the important problem of identifying which news is fake. However, because fake news datasets are often skewed, a high precision can be easily achieved by making fewer positive predictions. Thus, recall is used to measure the sensitivity, or the fraction of annotated fake news articles that are predicted to be fake news. F1 is used to combine precision and recall, which can provide an overall prediction performance for fake news detection. Note that for Precision, Recall, F1, and Accuracy, the higher the value, the better the performance.

The Receiver Operating Characteristics (ROC) curve provides a way of comparing the performance of classifiers by looking at the trade-off in the False Positive Rate (FPR) and the True Positive Rate (TPR). To draw the ROC curve, we plot the FPR on the x axis and TPR along the y axis. The ROC curve compares the performance of different classifiers by changing class distributions via a threshold. TPR and FPR are defined as follows

$$TPR = |TP|/(|TP|+|FN|)$$
$$FPR = |FP|/(FP|+|TN|)$$

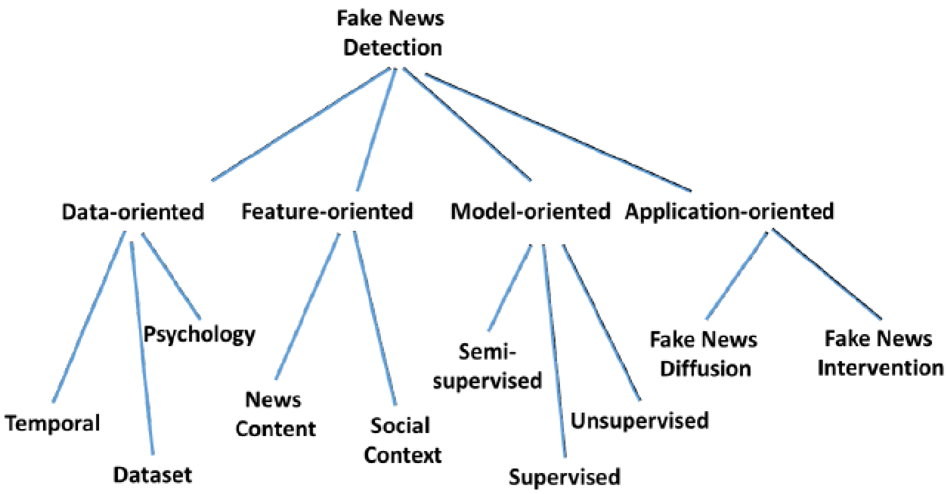

FIGURE 3: Future directions and open issues for fake news detection on social media.

Gap identified

In the existing system, news relating to US politics is correctly identified as fake or true. But except for that, most of the other news articles are wrongly identified as true or fake. That is why we are not using this system.

Proposed method

In this method, we are passing a set of data through NLP code and we are using the logistical regression model to process the data. We are using logistical regression because it is one of the simplest method to predict if an object is true or false (yes or no or 1 or 0). We have imported various libraries of python like pandas, NumPy, re, etc. We have also imported NLP libraries to help us process textual data. We have imported libraries to help us convert textual data into numeric data. We have also imported stopwords from English language to help us filter out stopwords for our prediction. Stopwords refer to those words which add no meaning to the data being processed.

We have loaded the file and then made specific modifications that assist us in our prediction. We have identified the columns which have null values and replaced them with an empty string. We have then merged two columns, namely the author column and the title column so that the prediction can be made easily. We have stored the combined column in “content” column.

The label column in the dataset has given the values 0 and 1 to the data depending on where the news is true or not. We have separated the label column from the rest of the dataset.

We have used the stemming process as one of the ways to filter out unnecessary and irrelevant characters from our dataset. Stemming refers to the process of reducing the words to its root word. For example, eats, eat and eating can be reduced to eat. We are also removing stopwords. Stopwords refer to those words which add no meaning to the data being processed. We are removing special characters like !, @, # ,$, etc from the text as these characters are irrelevant for our analysis. After making functions for stemming, stopwords and regular expression (!,@...) we apply them to the “contents” column.

After cleaning and pre-processing the data comes the next step of converting textual data into numeric data. As the computer cannot understand the textual language, we convert the text into numeric data using TfidfVectorizer(). TfidfVectorizer() counts the number of times a particular word is coming and it assigns a particular number to the words based on the frequency of its occurrence. It reduces the value of words repeating more and so feature vectors are created.

After this we divide the data into training dataset and test dataset. Training data builds the machine learning model. It teaches what the expected output looks like. A test data set is a data set that is independent of the training data set, but that follows the same probability distribution as the training data set.

We apply Logistic regression algorithm on the training dataset and try to find the accuracy of the dataset. From the dataset taken by us, the accuracy comes out to be approximately 0.98 which is quite good. We then applied the algorithm on the test dataset and the accuracy comes out to be approximately 0.97.

We made the predictive model to predict if the news article is fake or not. It gives the value 1 if it is fake and gives value 0 if it is true. We can then apply this algorithm and find out if the news is fake or not.

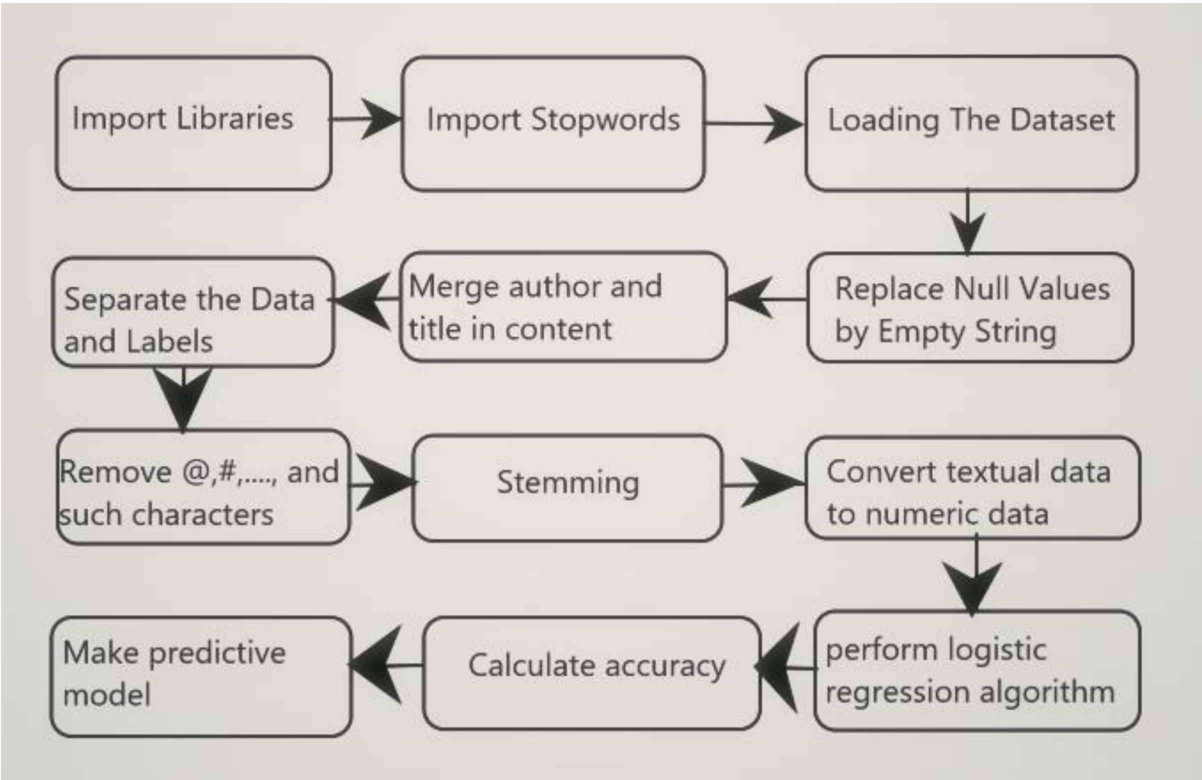

FIGURE 4: Flowchart showing the working of the model

Tools used: Google colab

Colab is a free notebook environment that runs entirely in the cloud. It lets you and your team members edit documents, the way you work with Google Docs. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.

This tutorial gives an exhaustive coverage of all the features of Colab and makes you comfortable working on it with confidence

Language used: python

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small- and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features such as list comprehensions, cycle-detecting garbage collection, reference counting, and Unicode support. Python 3.0, released in 2008, was a major revision that is not completely backward-compatible with earlier versions. Python 2 was discontinued with version 2.7.18 in 2020.

Python consistently ranks as one of the most popular programming languages.

Algorithm used

Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems.

Y=1/(1+e^-z)

Z=Wx+b

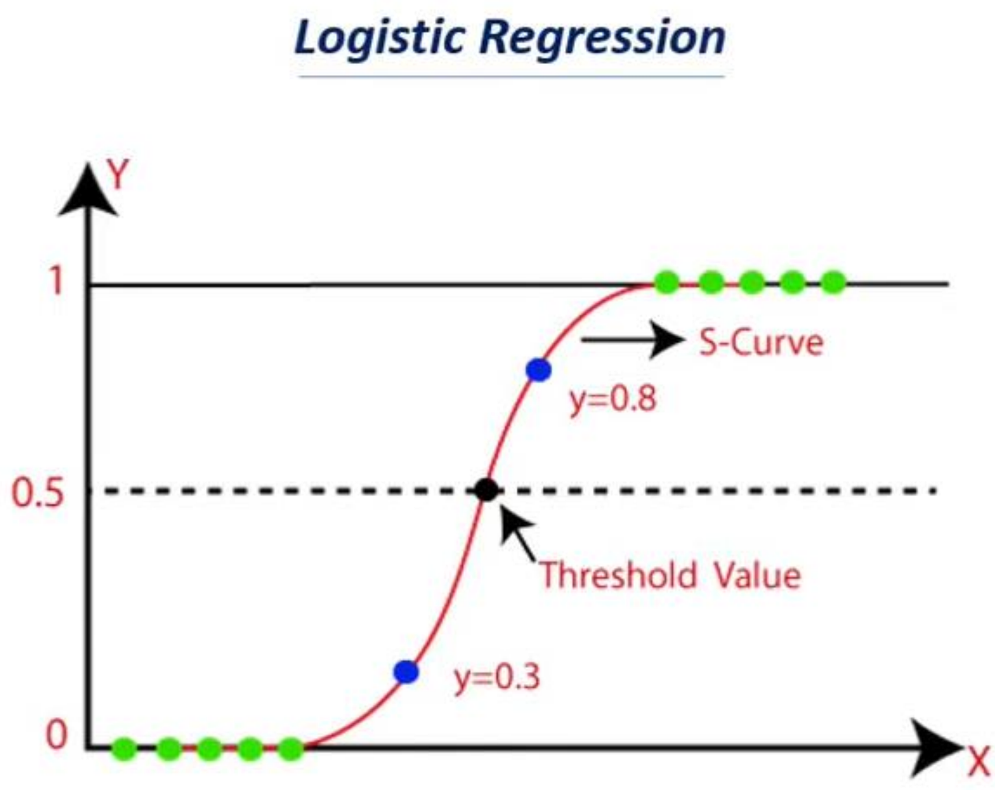

FIGURE 5: Sigmoid function for logical Regression

X: Input feature

Y: Prediction Probability

W: weight

B: baises

Code

```
import numpy as np
import pandas as pd
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score


# Fake -> 1
# Real -> 0
```

```python
import nltk
nltk.download('stopwords')


# printing the stopwords in English
print(stopwords.words('english'))


## Pre Processing of data
# loading the dataset to a pandas DataFrame
news_dataset = pd.read_csv('/content/train.csv')


#Number of rows and column in dataset
news_dataset.shape


# Print first 5 rows of the dataframe
news_dataset.head()

# Counting the number of missing values in the dataset
news_dataset.isnull()


# replacing the null values with empty string
news_dataset = news_dataset.fillna('')


# merging the author name and news title
news_dataset['content'] = news_dataset['author']+' '+news_dataset['titl
e']


print(news_dataset['content'])


# separating the data & label
X = news_dataset.drop(columns='label', axis=1)
Y = news_dataset['label']
print(X)
print(Y)


## Stemming
## Stemming is the process of reducing a word to its Root word
port_stem = PorterStemmer()


def stemming(content):
    stemmed_content = re.sub('[^a-zA-Z]',' ',content)
    stemmed_content = stemmed_content.lower()
```

```
    stemmed_content = stemmed_content.split()
    stemmed_content = [port_stem.stem(word) for word in stemmed_content
 if not word in stopwords.words('english')]
    stemmed_content = ' '.join(stemmed_content)
    return stemmed_content

news_dataset['content'] = news_dataset['content'].apply(stemming)

print(news_dataset['content'])

#separating the data and label
X = news_dataset['content'].values
Y = news_dataset['label'].values
print(X)
print(Y)
Y.shape

# converting the textual data to numerical data
vectorizer = TfidfVectorizer()
vectorizer.fit(X)
X = vectorizer.transform(X)

print(X)

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0
.2, stratify=Y, random_state=2)

## Training the model : Logistic Regression
model = LogisticRegression()
model.fit(X_train, Y_train)



## Evaluation
## Accuracy

# accuracy score on the training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

print('Accuracy score of the training data : ', training_data_accuracy)

# accuracy score on the test data
```

```
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)


print('Accuracy score of the test data : ', test_data_accuracy)


## Making a predictive model


X_new = X_test[3]

prediction = model.predict(X_new)
print(prediction)

if (prediction[0]==0):
  print('The news is Real')
else:
  print('The news is Fake')


print(Y_test[3])

print(Y_test[5])
```

Output

```
[1] import numpy as np
    import pandas as pd
    import re
    from nltk.corpus import stopwords
    from nltk.stem.porter import PorterStemmer
    from sklearn.feature_extraction.text import TfidfVectorizer
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score
```

```
[2] # Fake -> 1
    # Real -> 0
```

```
[3] import nltk
    nltk.download('stopwords')

    [nltk_data] Downloading package stopwords to /root/nltk_data...
    [nltk_data]   Unzipping corpora/stopwords.zip.
    True
```

```
[4] # printing the stopwords in English
    print(stopwords.words('english'))

    ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves',
```

```
[5] ## Pre Processing of data
```

```
[7] # loading the dataset to a pandas DataFrame
    news_dataset = pd.read_csv('/content/train.csv')
```

```
[8] #Number of rows and column in dataset
    news_dataset.shape

    (20800, 5)
```

```
[9] # Print first 5 rows of the dataframe
news_dataset.head()
```

	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

```
[10] # Counting the number of missing values in the dataset
news_dataset.isnull()
```

	id	title	author	text	label
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
...	...	...	...	...	...
20795	False	False	False	False	False
20796	False	False	False	False	False
20797	False	False	False	False	False
20798	False	False	False	False	False
20799	False	False	False	False	False

20800 rows × 5 columns

```
[11] # replacing the null values with empty string
news_dataset = news_dataset.fillna('')
```

```
[12] # merging the author name and news title
news_dataset['content'] = news_dataset['author']+' '+news_dataset['title']
```

```
[13] print(news_dataset['content'])
```

```
0        Darrell Lucus House Dem Aide: We Didn't Even S...
1        Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2        Consortiumnews.com Why the Truth Might Get You...
3        Jessica Purkiss 15 Civilians Killed In Single ...
4        Howard Portnoy Iranian woman jailed for fictio...
                              ...
20795    Jerome Hudson Rapper T.I.: Trump a 'Poster Chi...
20796    Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797    Michael J. de la Merced and Rachel Abrams Macy...
20798    Alex Ansary NATO, Russia To Hold Parallel Exer...
20799              David Swanson What Keeps the F-35 Alive
```

```
[14] # separating the data & label
X = news_dataset.drop(columns='label', axis=1)
Y = news_dataset['label']
```

```
[15] print(X)
```

```
          id                                              title  \
0          0  House Dem Aide: We Didn’t Even See Comey’s Let...
1          1  FLYNN: Hillary Clinton, Big Woman on Campus - ...
2          2                  Why the Truth Might Get You Fired
3          3  15 Civilians Killed In Single US Airstrike Hav...
4          4  Iranian woman jailed for fictional unpublished...
...      ...                                                ...
20795  20795  Rapper T.I.: Trump a ’Poster Child For White S...
20796  20796  N.F.L. Playoffs: Schedule, Matchups and Odds -...
20797  20797  Macy’s Is Said to Receive Takeover Approach by...
20798  20798  NATO, Russia To Hold Parallel Exercises In Bal...
20799  20799                          What Keeps the F-35 Alive

                                          author  \
0                                  Darrell Lucus
1                                Daniel J. Flynn
2                             Consortiumnews.com
3                                Jessica Purkiss
4                                 Howard Portnoy
...                                          ...
20795                              Jerome Hudson
20796                           Benjamin Hoffman
20797  Michael J. de la Merced and Rachel Abrams
20798                                Alex Ansary
20799                              David Swanson

                                                    text  \
0      House Dem Aide: We Didn’t Even See Comey’s Let...
1      Ever get the feeling your life circles the rou...
2      Why the Truth Might Get You Fired October 29, ...
3      Videos 15 Civilians Killed In Single US Airstr...
4      Print \nAn Iranian woman has been sentenced to...
...                                                  ...
20795  Rapper T. I. unloaded on black celebrities who...
20796  When the Green Bay Packers lost to the Washing...
20797  The Macy’s of today grew from the union of sev...
20798  NATO, Russia To Hold Parallel Exercises In Bal...
20799    David Swanson is an author, activist, journa...

                                                 content
0      Darrell Lucus House Dem Aide: We Didn’t Even S...
1      Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2      Consortiumnews.com Why the Truth Might Get You...
3      Jessica Purkiss 15 Civilians Killed In Single ...
4      Howard Portnoy Iranian woman jailed for fictio...
...                                                  ...
20795  Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796  Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797  Michael J. de la Merced and Rachel Abrams Macy...
20798  Alex Ansary NATO, Russia To Hold Parallel Exer...
20799            David Swanson What Keeps the F-35 Alive

[20800 rows x 5 columns]
```

```
[16] print(Y)
```

```
0        1
1        0
2        1
3        1
4        1
        ..
20795    0
20796    0
20797    0
20798    1
20799    1
Name: label, Length: 20800, dtype: int64
```

```
[17] ## Stemming
     ## Stemming is the process of reducing a word to its Root word
```

```
[18] port_stem = PorterStemmer()
```

```
[19] def stemming(content):
         stemmed_content = re.sub('[^a-zA-Z]',' ',content)
         stemmed_content = stemmed_content.lower()
         stemmed_content = stemmed_content.split()
         stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
         stemmed_content = ' '.join(stemmed_content)
         return stemmed_content
```

```
[20] news_dataset['content'] = news_dataset['content'].apply(stemming)
```

```
[21] print(news_dataset['content'])
```

```
0        darrel lucu hous dem aid even see comey letter...
1        daniel j flynn flynn hillari clinton big woman...
2                   consortiumnew com truth might get fire
3        jessica purkiss civilian kill singl us airstri...
4        howard portnoy iranian woman jail fiction unpu...
                               ...
20795    jerom hudson rapper trump poster child white s...
20796    benjamin hoffman n f l playoff schedul matchup...
20797    michael j de la merc rachel abram maci said re...
20798    alex ansari nato russia hold parallel exercis ...
20799                            david swanson keep f aliv
Name: content, Length: 20800, dtype: object
```

```
[22] #separating the data and label
     X = news_dataset['content'].values
     Y = news_dataset['label'].values
```

```
[23] print(X)
```

```
['darrel lucu hous dem aid even see comey letter jason chaffetz tweet'
 'daniel j flynn flynn hillari clinton big woman campu breitbart'
 'consortiumnew com truth might get fire' ...
 'michael j de la merc rachel abram maci said receiv takeov approach hudson bay new york time'
 'alex ansari nato russia hold parallel exercis balkan'
 'david swanson keep f aliv']
```

```
[24] print(Y)
```

```
[1 0 1 ... 0 1 1]
```

```
[25] Y.shape
```

```
(20800,)
```

```
[26] # converting the textual data to numerical data
     vectorizer = TfidfVectorizer()
     vectorizer.fit(X)

     X = vectorizer.transform(X)
```

```
print(X)
```

```
  (0, 15686)	0.28485063562728646
  (0, 13473)	0.2565896679337957
  (0, 8909)	0.3635963806326075
  (0, 8630)	0.29212514087043684
  (0, 7692)	0.24785219520671603
  (0, 7005)	0.21874169089359144
  (0, 4973)	0.233316966909351
  (0, 3792)	0.2705332480845492
  (0, 3600)	0.3598939188262559
  (0, 2959)	0.2468450128533713
  (0, 2483)	0.3676519686797209
  (0, 267)	0.27010124977708766
  (1, 16799)	0.30071745655510157
  (1, 6816)	0.1904660198296849
  (1, 5503)	0.7143299355715573
  (1, 3568)	0.26373768806048464
  (1, 2813)	0.19094574062359204
  (1, 2223)	0.3827320386859759
  (1, 1894)	0.15521974226349364
  (1, 1497)	0.2939891562094648
  (2, 15611)	0.41544962664721613
  (2, 9620)	0.49351492943649944
  (2, 5968)	0.3474613386728292
  (2, 5389)	0.3866530551182615
  (2, 3103)	0.46097489583229645
  :	:
  (20797, 13122)	0.2482526352197606
  (20797, 12344)	0.27263457663336677
  (20797, 12138)	0.24778257724396507
  (20797, 10306)	0.08038079000566466
  (20797, 9588)	0.174553480255222
  (20797, 9518)	0.2954204003420313
  (20797, 8988)	0.36160868928090795
  (20797, 8364)	0.22322585870464118
  (20797, 7042)	0.21799048897828688
  (20797, 3643)	0.21155500613623743
  (20797, 1287)	0.33538056804139865
  (20797, 699)	0.30685846079762347
  (20797, 43)	0.29710241860700626
  (20798, 13046)	0.22363267488270608
  (20798, 11052)	0.4460515589182236
  (20798, 10177)	0.3192496370187028
  (20798, 6889)	0.32496285694299426
  (20798, 5032)	0.4083701450239529
  (20798, 1125)	0.4460515589182236
  (20798, 588)	0.3112141524638974
  (20798, 350)	0.28446937819072576
  (20799, 14852)	0.5677577267055112
  (20799, 8036)	0.45983893273780013
  (20799, 3623)	0.37927626273066584
  (20799, 377)	0.5677577267055112
```

```
[28] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, stratify=Y, random_state=2)
```

```
[29] ## Training the model : Logistic Regression
```

```
[30] model = LogisticRegression()
```

```
[31] model.fit(X_train, Y_train)
```

```
LogisticRegression()
```

```
[32] ## Evaluation
     ## Accuracy
```

```
[33] # accuracy score on the training data
     X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
[34] print('Accuracy score of the training data : ', training_data_accuracy)
```

```
Accuracy score of the training data :  0.9865985576923076
```

```
[35] # accuracy score on the test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[36] print('Accuracy score of the test data : ', test_data_accuracy)

     Accuracy score of the test data :  0.9790865384615385

[37] ## Making a predictive model

[38] X_new = X_test[3]

     prediction = model.predict(X_new)
     print(prediction)

     if (prediction[0]==0):
       print('The news is Real')
     else:
       print('The news is Fake')

     [0]
     The news is Real

[39] print(Y_test[3])

     0

[40] print(Y_test[5])

     1
```

FIGURE 6: Output got from the proposed model

Sample data

https://www.kaggle.com/c/fake-news/data?select=train.csv

Conclusion

With the increasing popularity of social media, more and more people consume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which has strong negative impacts on individual users and broader society.

I have presented in this project some revealing characteristics about fake news. We have seen how the news is pre-processed and made ready for the analysis. We also saw how logistical regression can be used to predict if the news articles were fake or true.

References

[1] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In ISSP'12.

[2] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[3] Solomon E Asch and H Guetzkow. Effects of group pressure upon the modification and distortion of judgments. Groups, leadership, and men, pages 222–236, 1951.

[4] Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. Communication Research, 41(3):430–454, 2014.

[5] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In IJCAI'07.

[6] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. First Monday, 21(11), 2016.

[7] Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. " 8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In AAAI'16.

[8] Jonas Nygaard Blom and Kenneth Reinecke Hansen. Click bait: Forward-reference as lure in online news headlines. Journal of Pragmatics, 76:87–100, 2015.

[9] Paul R Brewer, Dannagal Goldthwaite Young, and Michelle Morreale. The impact of real news about fake news: Intertextual processes and political satire. International Journal of Public Opinion Research, 25(3):323–343, 2013.

[10] Carlos Castillo, Mohammed El-Haddad, J¨urgen Pfeffer, and Matt Stempeck. Characterizing the life cycle of online news stories using social media reactions. In CSCW'14.

[11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In WWW'11.

[12] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In ASONAM'16.

[13] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, pages 15–19. ACM, 2015.

[14] Justin Cheng, Michael Bernstein, Cristian DanescuNiculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In CSCW '17.

[15] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing, 9(6):811–824, 2012.