Project Report on

SPAM NEWS DETECTION

- Done By

L VIKRAM SIMHA REDDY

Table of Content

1. Project Definition

In the technological era, the propagation of fake news has come to be a huge irritation, with public opinion being the worst victim most of the time, transmitting misinformation, and even immigrating political ecosystems. The goal of this project is to build a fake news detection system that can be used flexibly and that is stably operational.

The primary aim of it is to compose a neat differentiate fake and real news computer program. To do this, we will have to gather a really huge dataset of both news articles real and bogus, and the next step will be to train a machine learning model by using this dataset. By analyzing linguistic patterns, word usages, source credibility, and metadata, the model is anticipated to make the right decision on which news is trustworthy and which is an imposter. Apart from that, the project will experiment with different NLP techniques like sentiment analysis, topic modeling, and text classification to improve the detection rate.

Thus, getting instantaneous feedback on whether an article is fake or not will make users become alert in the way they process information and pass it to others.

2. Literature Survey

Rubin searched for good ways to use linguistic hints and machine learning to identify fake news. They said that the primary thing is the story and the style so, through it, the real content is discriminated. Besides, they also proposed a model hardware, which is a composite of linguistics to identify information that is misleading or wrong.

Conroy gave a thorough overview of automatic cheating detection technology. They did a comparative study of different machine learning methods, such as Support Vector Machines (SVM) and Naive Bayes and how well they could find fake news. They also pointed out the barriers of deception in short texts and the need for more sophisticated algorithms.

Volkova examined the rising effect of social media on fake news dissemination. They studied the trends of the spread and introduced a solution that relies on user's behavior and network analysis for the detection of misinformation. Their study suggested that social clues are necessary to improve detection systems.

Shu have presented a model using fake news detection that is comprised of content, social domain, and credibility. They used models like machine learning along with network-based methods for the better detection of fake new. Their research underlined the fact that a multi-dimensional approach is essential for the detection of fake new.

All together, these studies carry stress to the dearth of combining different techniques primarily NLP, machine learning, and network analysis, to gain the accuracy and robustness of the fake news detection systems. The always changing nature of the very first news-making way requires that there be more and more the effectiveness, and these only should suffice to be all.

3. Python & Jupyter Notebook

Python is a high-level, interpreted programming language known for its readability and versatility. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python's extensive standard library and a vast ecosystem of third-party packages make it a popular choice for web development, data analysis, artificial intelligence, scientific computing, and more.

Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text. It supports many programming languages, with Python being the most common. Jupyter Notebooks enable users to write code and see the results immediately. This interactivity makes it an excellent tool for data analysis, visualization, and exploratory programming. Jupyter Notebook provides an interactive platform for coding, data analysis, and documentation, making it a valuable tool for many Python developers and data scientists.

Installing Jupyter Notebook Using pip

To install Jupyter Notebook using pip, the package installer for Python, follow these steps:

1. **Ensure Python is Installed:** First, check if Python is installed on your system. You can do this by running the following command in your terminal or command prompt:

python --version

If Python is not installed, download and install the latest version from official python website.

2. **Install Jupyter Notebook:** Use pip to install Jupyter Notebook by running the following command:

pip install notebook

This command will install Jupyter Notebook along with all its dependencies.

3. **Launching Jupyter Notebook:** After installation, you can launch Jupyter Notebook by running:

jupyter notebook

This command will open the Jupyter Notebook interface in your default web browser. You can create new notebooks, open existing ones, and start writing and executing code.

4. Python Packages and their descriptions

NumPy (Numerical Python) is a fundamental package for scientific computing in Python. It provides support for arrays, which are collections of elements (usually numbers), and a vast array of mathematical functions to operate on these arrays. NumPy's array object, `ndarray`, allows for efficient storage and manipulation of large datasets, making it essential for numerical computations. Key features include:

- Support for multi-dimensional arrays and matrices.
- Mathematical functions for linear algebra, and random number generation.

Pandas is a powerful open-source data analysis and manipulation library built on top of NumPy. It provides data structures like `DataFrame` and `Series` that make it easy to handle structured data, such as tabular data, time series, and other forms of data. Key features include:

- `DataFrame`: A two-dimensional, size-mutable, and heterogeneous data structure with labeled axes (rows and columns).
- `Series`: A one-dimensional labeled array capable of holding any data type.
- Data alignment and integrated handling of missing data.
- Tools for reading and writing data in various formats, such as CSV, Excel, SQL databases, and HDF5.

Scikit-learn is a comprehensive machine learning library in Python. It provides simple and efficient tools for data mining and data analysis, built on NumPy, SciPy, and matplotlib. Scikit-learn is widely used for implementing and deploying machine learning algorithms. Key features include:

- A wide range of supervised and unsupervised learning algorithms, such as linear regression, support vector machines, decision trees, clustering, and more.
- Tools for model selection, validation, and evaluation.
- Preprocessing tools for feature extraction and normalization.

NLTK (Natural Language Toolkit) is a leading platform for building Python programs to work with human language data (text). It provides a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and more. Key features include:

- Access to a variety of corpora and lexical resources, such as WordNet.
- Text processing modules for tokenization, stemming, and tagging.
- Tools for syntactic parsing, semantic interpretation, and machine learning.
- A suite of libraries for building and evaluating language models.

Re is a module in Python that provides support for regular expressions, a powerful tool for pattern matching and text manipulation. Regular expressions are used to identify specific patterns within strings, such as validating email addresses, extracting substrings, or replacing text. Key features include:

- Support for complex pattern matching using metacharacters and special sequences.
- Integration with Python's string methods for seamless text manipulation.

These packages are widely used in various fields, including data science, machine learning, natural language processing, and software development, due to their efficiency, ease of use, and extensive functionality.

5. Installing Python Libraries in Jupyter Notebook

To install the specified Python packages, you can use the pip package manager. Here are the installation commands for each package:

NumPy:

pip install numpy

pandas:

pip install pandas

NLTK:

pip install nltk

re: The re module is part of Python's standard library, so it does not require a separate installation. It is included with Python and can be imported directly.

stopwords and WordNetLemmatizer (via NLTK): The stopwords and WordNetLemmatizer are resources within the NLTK library. After installing NLTK, you can download these resources using the following Python commands:

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

Alternatively, you can download these resources in a script or Jupyter Notebook by adding these lines of code after importing NLTK.

These commands will install the necessary packages and resources, allowing you to use them in your Python projects.

6. Python Code for Implementation

spam_news_detection

August 3, 2024

[1]:
```
import numpy as np
import pandas as pd
```

[2]:
```
true_news = pd.read_csv("True.csv")
fake_news = pd.read_csv("Fake.csv")
```

[3]:
```
true_news
```

[3]:

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t…	WASHINGTON (Reuters) - The head of a conservat…	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o…	WASHINGTON (Reuters) - Transgender people will…	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell…	WASHINGTON (Reuters) - The special counsel inv…	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat…	WASHINGTON (Reuters) - Trump campaign adviser …	politicsNews	
4	Trump wants Postal Service to charge 'much mor…	SEATTLE/WASHINGTON (Reuters) - President Donal…	politicsNews	
…	…	…	…	
21412	'Fully committed' NATO backs new U.S. approach…	BRUSSELS (Reuters) - NATO allies on Tuesday we…	worldnews	
21413	LexisNexis withdrew two products from Chinese …	LONDON (Reuters) - LexisNexis, a provider of l…	worldnews	
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov…	worldnews	
21415	Vatican upbeat on possibility of Pope Francis …	MOSCOW (Reuters) - Vatican Secretary of State …	worldnews	
21416	Indonesia to buy $1.14 billion worth of Russia…	JAKARTA (Reuters) - Indonesia will buy 11 Sukh…	worldnews	

```
3       December 30, 2017
4       December 29, 2017
…                     …
21412     August 22, 2017
21413     August 22, 2017
21414     August 22, 2017
21415     August 22, 2017
21416     August 22, 2017

[21417 rows x 4 columns]
```

[4]:
```
fake_news
```

[4]:
```
                                                   title  \
0        Donald Trump Sends Out Embarrassing New Year'…
1        Drunk Bragging Trump Staffer Started Russian …
2        Sheriff David Clarke Becomes An Internet Joke…
3        Trump Is So Obsessed He Even Has Obama's Name…
4        Pope Francis Just Called Out Donald Trump Dur…
…                                                    …
23476  McPain: John McCain Furious That Iran Treated …
23477  JUSTICE? Yahoo Settles E-mail Privacy Class-ac…
23478  Sunnistan: US and Allied 'Safe Zone' Plan to T…
23479  How to Blow $700 Million: Al Jazeera America F…
23480  10 U.S. Navy Sailors Held by Iranian Military …

                                                    text       subject  \
0      Donald Trump just couldn t wish all Americans …          News
1      House Intelligence Committee Chairman Devin Nu…          News
2      On Friday, it was revealed that former Milwauk…          News
3      On Christmas day, Donald Trump announced that …          News
4      Pope Francis used his annual Christmas Day mes…          News
…                                                    …         …
23476  21st Century Wire says As 21WIRE reported earl…  Middle-east
23477  21st Century Wire says It s a familiar theme. …  Middle-east
23478  Patrick Henningsen  21st Century WireRemember …  Middle-east
23479  21st Century Wire says Al Jazeera America will…  Middle-east
23480  21st Century Wire says As 21WIRE predicted in …  Middle-east

                   date
0      December 31, 2017
1      December 31, 2017
2      December 30, 2017
3      December 29, 2017
4      December 25, 2017
…                     …
23476   January 16, 2016
```

```
23477   January 16, 2016
23478   January 15, 2016
23479   January 14, 2016
23480   January 12, 2016

[23481 rows x 4 columns]
```

[5]:
```
true_news['label'] = 0
fake_news['label'] = 1
```

[6]:
```
true_news
```

[6]:
```
                                                   title  \
0      As U.S. budget fight looms, Republicans flip t…
1      U.S. military to accept transgender recruits o…
2      Senior U.S. Republican senator: 'Let Mr. Muell…
3      FBI Russia probe helped by Australian diplomat…
4      Trump wants Postal Service to charge 'much mor…
…                                                    …
21412  'Fully committed' NATO backs new U.S. approach…
21413  LexisNexis withdrew two products from Chinese …
21414  Minsk cultural hub becomes haven from authorities
21415  Vatican upbeat on possibility of Pope Francis …
21416  Indonesia to buy $1.14 billion worth of Russia…

                                                    text       subject  \
0      WASHINGTON (Reuters) - The head of a conservat…  politicsNews
1      WASHINGTON (Reuters) - Transgender people will…  politicsNews
2      WASHINGTON (Reuters) - The special counsel inv…  politicsNews
3      WASHINGTON (Reuters) - Trump campaign adviser …  politicsNews
4      SEATTLE/WASHINGTON (Reuters) - President Donal…  politicsNews
…                                                    …           …
21412  BRUSSELS (Reuters) - NATO allies on Tuesday we…     worldnews
21413  LONDON (Reuters) - LexisNexis, a provider of l…     worldnews
21414  MINSK (Reuters) - In the shadow of disused Sov…     worldnews
21415  MOSCOW (Reuters) - Vatican Secretary of State …     worldnews
21416  JAKARTA (Reuters) - Indonesia will buy 11 Sukh…     worldnews

                    date  label
0      December 31, 2017       0
1      December 29, 2017       0
2      December 31, 2017       0
3      December 30, 2017       0
4      December 29, 2017       0
…                    …       …
21412    August 22, 2017       0
21413    August 22, 2017       0
```

```
21414    August 22, 2017       0
21415    August 22, 2017       0
21416    August 22, 2017       0

[21417 rows x 5 columns]
```

[7]:
```
fake_news
```

[7]:
```
                                                   title  \
0       Donald Trump Sends Out Embarrassing New Year'…
1       Drunk Bragging Trump Staffer Started Russian …
2       Sheriff David Clarke Becomes An Internet Joke…
3       Trump Is So Obsessed He Even Has Obama's Name…
4       Pope Francis Just Called Out Donald Trump Dur…
…                                                     …
23476  McPain: John McCain Furious That Iran Treated …
23477  JUSTICE? Yahoo Settles E-mail Privacy Class-ac…
23478  Sunnistan: US and Allied 'Safe Zone' Plan to T…
23479  How to Blow $700 Million: Al Jazeera America F…
23480  10 U.S. Navy Sailors Held by Iranian Military …

                                                    text       subject  \
0      Donald Trump just couldn t wish all Americans …          News
1      House Intelligence Committee Chairman Devin Nu…          News
2      On Friday, it was revealed that former Milwauk…          News
3      On Christmas day, Donald Trump announced that …          News
4      Pope Francis used his annual Christmas Day mes…          News
…                                                    …           …
23476  21st Century Wire says As 21WIRE reported earl…  Middle-east
23477  21st Century Wire says It s a familiar theme. …  Middle-east
23478  Patrick Henningsen  21st Century WireRemember …  Middle-east
23479  21st Century Wire says Al Jazeera America will…  Middle-east
23480  21st Century Wire says As 21WIRE predicted in …  Middle-east

                    date  label
0       December 31, 2017      1
1       December 31, 2017      1
2       December 30, 2017      1
3       December 29, 2017      1
4       December 25, 2017      1
…                     …      …
23476    January 16, 2016      1
23477    January 16, 2016      1
23478    January 15, 2016      1
23479    January 14, 2016      1
23480    January 12, 2016      1
```

```
[23481 rows x 5 columns]
```

[8]:
```
dataset1 = true_news[['text','label']]
dataset2 = fake_news[['text','label']]
```

[9]:
```
dataset1
```

[9]:

	text	label
0	WASHINGTON (Reuters) - The head of a conservat…	0
1	WASHINGTON (Reuters) - Transgender people will…	0
2	WASHINGTON (Reuters) - The special counsel inv…	0
3	WASHINGTON (Reuters) - Trump campaign adviser …	0
4	SEATTLE/WASHINGTON (Reuters) - President Donal…	0
…	…	…
21412	BRUSSELS (Reuters) - NATO allies on Tuesday we…	0
21413	LONDON (Reuters) - LexisNexis, a provider of l…	0
21414	MINSK (Reuters) - In the shadow of disused Sov…	0
21415	MOSCOW (Reuters) - Vatican Secretary of State …	0
21416	JAKARTA (Reuters) - Indonesia will buy 11 Sukh…	0

```
[21417 rows x 2 columns]
```

[10]:
```
dataset2
```

[10]:

	text	label
0	Donald Trump just couldn t wish all Americans …	1
1	House Intelligence Committee Chairman Devin Nu…	1
2	On Friday, it was revealed that former Milwauk…	1
3	On Christmas day, Donald Trump announced that …	1
4	Pope Francis used his annual Christmas Day mes…	1
…	…	…
23476	21st Century Wire says As 21WIRE reported earl…	1
23477	21st Century Wire says It s a familiar theme. …	1
23478	Patrick Henningsen 21st Century WireRemember …	1
23479	21st Century Wire says Al Jazeera America will…	1
23480	21st Century Wire says As 21WIRE predicted in …	1

```
[23481 rows x 2 columns]
```

[11]:
```
dataset = pd.concat([dataset1, dataset2])
```

[12]:
```
dataset
```

[12]:

	text	label
0	WASHINGTON (Reuters) - The head of a conservat…	0
1	WASHINGTON (Reuters) - Transgender people will…	0
2	WASHINGTON (Reuters) - The special counsel inv…	0
3	WASHINGTON (Reuters) - Trump campaign adviser …	0

```
4      SEATTLE/WASHINGTON (Reuters) - President Donal…      0
…                                                    …    …
23476  21st Century Wire says As 21WIRE reported earl…      1
23477  21st Century Wire says It s a familiar theme. …      1
23478  Patrick Henningsen  21st Century WireRemember …      1
23479  21st Century Wire says Al Jazeera America will…      1
23480  21st Century Wire says As 21WIRE predicted in …      1

[44898 rows x 2 columns]
```

[13]:
```
dataset.shape
```

[13]:
```
(44898, 2)
```

[14]:
```
dataset.isnull().sum()
```

[14]:
```
text     0
label    0
dtype: int64
```

[15]:
```
dataset['label'].value_counts()
```

[15]:
```
label
1    23481
0    21417
Name: count, dtype: int64
```

[16]:
```
dataset = dataset.sample(frac = 1)
```

[17]:
```
dataset
```

[17]:
```
                                                    text  label
16857  BERLIN (Reuters) - German Chancellor Angela Me…      0
15209  BUENOS AIRES (Reuters) - A Russian official ex…      0
7562   Marco Rubio has officially joined the anti-Tru…      1
19449  Politico has terminated its contract with maga…      1
13283  Don t mess with daddy s little girl! Ivanka Tr…      1
…                                                    …    …
2513   PARIS (Reuters) - France's foreign ministry on…      0
18755  Now one unlikely Senator is about to put forth…      1
18380  Beyond the pale ABC reporter Matthew Dowd comp…      1
3329   President-elect Donald Trump loves bragging ab…      1
17020  WASHINGTON (Reuters) - U.S. Secretary of State…      0

[44898 rows x 2 columns]
```

[18]:
```
import nltk
import re
```

```
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
```

[19]:
```
ps  = WordNetLemmatizer()
```

[20]:
```
nltk.download('wordnet')
nltk.download('omw-1.4')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package wordnet to C:\Users\VENKATA SAI
[nltk_data]     RAM\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to C:\Users\VENKATA SAI
[nltk_data]     RAM\AppData\Roaming\nltk_data…
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package stopwords to C:\Users\VENKATA SAI
[nltk_data]     RAM\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

[20]: True

[21]:
```
stopwords = stopwords.words('english')
```

[22]:
```
def clean_row(row):
    row = row.lower()
    row = re.sub('[^a-zA-Z]', ' ', row)
    token = row.split()
    news = [ps.lemmatize(word) for word in token if not word in stopwords]
    cleaned_news = ' '.join(news)
    return cleaned_news
```

[23]:
```
dataset['text']
```

[23]:
```
16857    BERLIN (Reuters) - German Chancellor Angela Me…
15209    BUENOS AIRES (Reuters) - A Russian official ex…
7562     Marco Rubio has officially joined the anti-Tru…
19449    Politico has terminated its contract with maga…
13283    Don t mess with daddy s little girl! Ivanka Tr…
                              …
2513     PARIS (Reuters) - France's foreign ministry on…
18755    Now one unlikely Senator is about to put forth…
18380    Beyond the pale ABC reporter Matthew Dowd comp…
3329     President-elect Donald Trump loves bragging ab…
17020    WASHINGTON (Reuters) - U.S. Secretary of State…
Name: text, Length: 44898, dtype: object
```

[24]:
```
dataset['text'] = dataset['text'].apply(lambda x : clean_row(x))
```

```
[25]: dataset['text']
```

```
[25]: 16857    berlin reuters german chancellor angela merkel…
      15209    buenos aire reuters russian official expressed…
      7562     marco rubio officially joined anti trump squad…
      19449    politico terminated contract magazine writer j…
      13283    mess daddy little girl ivanka trump nobody foo…
                                    …
      2513     paris reuters france foreign ministry wednesda…
      18755    one unlikely senator put forth bill paving way…
      18380    beyond pale abc reporter matthew dowd compared…
      3329     president elect donald trump love bragging two…
      17020    washington reuters u secretary state rex tille…
      Name: text, Length: 44898, dtype: object
```

```
[26]: from sklearn.feature_extraction.text import TfidfVectorizer
```

```
[27]: vectorizer = TfidfVectorizer(max_features = 50000, lowercase = False,␣
       ↪ngram_range = (1, 2))
```

```
[28]: x = dataset.iloc[:40000, 0]
      y = dataset.iloc[:40000, 1]
```

```
[29]: x
```

```
[29]: 16857    berlin reuters german chancellor angela merkel…
      15209    buenos aire reuters russian official expressed…
      7562     marco rubio officially joined anti trump squad…
      19449    politico terminated contract magazine writer j…
      13283    mess daddy little girl ivanka trump nobody foo…
                                    …
      4128     today one national poll clinton leading male v…
      16344    former u attorney joseph digenova slammed fbi …
      5248     khizr khan whose son killed iraq serving unite…
      5946     anyone clearly enough donald trump ridiculous …
      4247     washington reuters president donald j trump pl…
      Name: text, Length: 40000, dtype: object
```

```
[30]: y
```

```
[30]: 16857    0
      15209    0
      7562     1
      19449    1
      13283    1
              ..
      4128     1
      16344    1
```

```
5248     1
5946     1
4247     0
Name: label, Length: 40000, dtype: int64
```

[31]:
```
from sklearn.model_selection import train_test_split
```

[32]:
```
train_data, test_data, train_label, test_label = train_test_split(x, y,␣
↪test_size = 0.2, random_state = 0)
```

[33]:
```
vec_train_data = vectorizer.fit_transform(train_data)
```

[34]:
```
vec_train_data = vec_train_data.toarray()
```

[35]:
```
vec_test_data = vectorizer.fit_transform(test_data)
```

[36]:
```
vec_test_data = vec_test_data.toarray()
```

[37]:
```
vec_train_data.shape, vec_test_data.shape
```

[37]:
```
((32000, 50000), (8000, 50000))
```

[38]:
```
train_data = pd.DataFrame(vec_train_data, columns = vectorizer.
↪get_feature_names_out())
test_data = pd.DataFrame(vec_test_data, columns = vectorizer.
↪get_feature_names_out())
```

[39]:
```
train_data
```

[39]:

	aa	aapl	aaron	aaron burr	aarp	ab	aback	abadi	abadi said	\
0	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	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
…	…	…	…	…	…	…	…	…	…	
31995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
31996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
31997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
31998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
31999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

	abandon	…	zone would	zoo	zor	zor province	zschaepe	zucker	\
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	
2	0.0	…	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	…	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	…	0.0	0.0	0.0	0.0	0.0	0.0	

```
…       …  …          …  …  …              …       …     …
31995     0.0  …         0.0  0.0  0.0          0.0     0.0   0.0
31996     0.0  …         0.0  0.0  0.0          0.0     0.0   0.0
31997     0.0  …         0.0  0.0  0.0          0.0     0.0   0.0
31998     0.0  …         0.0  0.0  0.0          0.0     0.0   0.0
31999     0.0  …         0.0  0.0  0.0          0.0     0.0   0.0

       zuckerberg  zulia  zuma  zurich
0             0.0    0.0   0.0     0.0
1             0.0    0.0   0.0     0.0
2             0.0    0.0   0.0     0.0
3             0.0    0.0   0.0     0.0
4             0.0    0.0   0.0     0.0
…           …    …   …     …
31995         0.0    0.0   0.0     0.0
31996         0.0    0.0   0.0     0.0
31997         0.0    0.0   0.0     0.0
31998         0.0    0.0   0.0     0.0
31999         0.0    0.0   0.0     0.0

[32000 rows x 50000 columns]
```

[40]:
```
test_data
```

[40]:
```
        aa  aapl  aaron  aaron burr      aarp   ab  aback  abadi  abadi said  \
0      0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0
1      0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0
2      0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0
3      0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0
4      0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0
…    …   …    …         …       …  …    …    …         …
7995   0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0
7996   0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0
7997   0.0   0.0    0.0         0.0  0.116119  0.0    0.0    0.0         0.0
7998   0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0
7999   0.0   0.0    0.0         0.0  0.000000  0.0    0.0    0.0         0.0

       abandon  …  zone would  zoo  zor  zor province  zschaepe  zucker  \
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
2          0.0  …         0.0  0.0  0.0           0.0       0.0     0.0
3          0.0  …         0.0  0.0  0.0           0.0       0.0     0.0
4          0.0  …         0.0  0.0  0.0           0.0       0.0     0.0
…        …  …         …  …  …             …       …     …
7995       0.0  …         0.0  0.0  0.0           0.0       0.0     0.0
7996       0.0  …         0.0  0.0  0.0           0.0       0.0     0.0
7997       0.0  …         0.0  0.0  0.0           0.0       0.0     0.0
```

```
7998      0.0  …         0.0  0.0  0.0          0.0       0.0      0.0
7999      0.0  …         0.0  0.0  0.0          0.0       0.0      0.0

      zuckerberg  zulia  zuma  zurich
0            0.0    0.0   0.0     0.0
1            0.0    0.0   0.0     0.0
2            0.0    0.0   0.0     0.0
3            0.0    0.0   0.0     0.0
4            0.0    0.0   0.0     0.0
…            …      …     …       …
7995         0.0    0.0   0.0     0.0
7996         0.0    0.0   0.0     0.0
7997         0.0    0.0   0.0     0.0
7998         0.0    0.0   0.0     0.0
7999         0.0    0.0   0.0     0.0

[8000 rows x 50000 columns]
```

[41]:
```
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVC
```

[42]:
```
clf = MultinomialNB()
#clf = LogisticRegression()
#clf = LinearRegression()
#clf = SVC()
```

[43]:
```
clf.fit(train_data, train_label)
```

[43]:
```
MultinomialNB()
```

[44]:
```
y_pred = clf.predict(test_data)
```

[45]:
```
test_label
```

[45]:
```
19765    0
15449    1
13137    1
12710    1
9954     1
        ..
14199    1
9028     0
10254    1
5940     1
15714    1
```

```
Name: label, Length: 8000, dtype: int64
```

[46]:
```
y_pred
```

[46]:
```
array([1, 1, 0, …, 1, 1, 0], dtype=int64)
```

[47]:
```
from sklearn.metrics import accuracy_score
```

[48]:
```
accuracy_score(test_label, y_pred)
```

[48]:
```
0.729625
```

[49]:
```
y_pred_train = clf.predict(train_data)
```

[50]:
```
accuracy_score(train_label, y_pred_train)
```

[50]:
```
0.95959375
```

[52]:
```
# SL>NO              METHOD              ACCURACY[TEST]
#    1           naive_bayes                  72%
#    2        logistic_regression             53%
```

[54]:
```
txt = input("Enter News")
news = clean_row(str(txt))
pred = clf.predict(vectorizer.transform([news]).toarray())
```

```
Enter News On Christmas day, Donald Trump announced that he would  be back to
work  the following day, but he is golfing for the fourth day in a row. The
former reality show star blasted former President Barack Obama for playing golf
and now Trump is on track to outpace the number of golf games his predecessor
played.Updated my tracker of Trump s appearances at Trump properties.71 rounds
of 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.twitter.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
golf:  Unlike Obama, we are working to fix the problem   and not on the golf
course.  However, the coding wasn t done correctly.The website of Donald Trump,
who has spent several days in a row at the golf course, is coded to serve up the
following message in 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/ayBlGmk65Z. 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. Still not clear when these messages would
actually display, since the actual 404 (and presumably 500) page displays a
different message pic.twitter.com/Z7dmyQ5smy  Christopher Ingraham (@_cingraham)
December 29, 2017That suggests someone at either RNC or the 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 code
was f-cked up.The best part about this is that they are using the  =
(assignment) operator which means that bit of code will never get run. If you
look a few lines up  errorCode  will always be  404          (@tw1trsux)
December 28, 2017trump s coders 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.

C:\Users\VENKATA SAI RAM\AppData\Local\Programs\Python\Python312\Lib\site-
packages\sklearn\base.py:493: UserWarning: X does not have valid feature names,
but MultinomialNB was fitted with feature names
  warnings.warn(
```

[55]:
```
if pred == 0:
  print('News is True')
else:
  print('News is Fake')
```

```
News is Fake
```

7. Logistic Regression and Naive Bayes(MultinomialNB)

Logistic Regression and **Naive Bayes** are two classifications widely applied to machine learning. Their core principles are quite different, and the applications are also well diversified.

Logistic Regression is a linear model for binary classification. It predicts a binary outcome, which generally includes 0 and 1 or even True and False from one or more input features. Logistic regression model the probability that a given input belongs to a particular class using a logistic function (or a sigmoid function). The sigmoid function satisfies the expectation of the minimum expected value and linear components. Therefore, the claimed routine can easily be transformed to a value between 0 and 1, which is actually the probability of a given event happening. Logistic regression method comes up with a decision on a new assessed score (commonly 0.5). Further, Logistic regression is an interpretable method; a model's coefficients, in simple terms, represent the change in log-odds of the outcome given the one unit increase in the predictor variable, which is suitable for such areas as medical diagnosis and financial risk assessment.

Naive Bayes is a stochastic classifier founded on Bayes' theorem by the "naive" supposition that features are independent of each other when conditioned on the class label. This approach documents the posterior probabilities, which in turn facilitate the more rapid classification, especially when working with high-dimensional datasets. Naive Bayes exists in different forms, such as Gaussian Naive Bayes (for continuous features), Multinomial Naive Bayes (for count data), and Bernoulli Naive Bayes (for binary features). The model achieves the posterior probabilities of all the categories and then assigns the one whose probability is the highest to the given input. Despite its strong independence assumption, Naive Bayes is an excellent performer in real-world situations, especially in the text categorization tasks, as well, such as the fields of filtering spams and sentiment analysis. Great deal of accuracy and applicability makes it a vital help to nlp engineering, as well as to numerous other applications.

8. Result

The result of the spam news detection model trained using nlp and machine-learning using the methods Logistic Regression and Naive Bayes(MultinomialNB) is tabulated below.

S.No	Method	Accuracy
1	*naive_bayes(MultinomialNB)*	*72%*
2	*Logistic regression*	*53%*

Out of the two methods, naive_bayes(MultinomialNB) has accuracy 72% which is far better than logistic regression having accuracy 53%.

9. Conclusion

The **Spam News Detection Project** aims to distinguish between legitimate news and spam or fake news using machine learning techniques. The project utilizes a combination of key Python libraries, including **NumPy** for numerical computations, **pandas** for data manipulation, **NLTK** for natural language processing, and **scikit-learn** for implementing machine learning algorithms such as Logistic Regression and Naive Bayes.

The dataset comprises news articles labelled as either legitimate or spam. The project involves preprocessing the text data, including tokenization, stopwords removal, and lemmatization, using NLTK. Feature extraction methods, such as term frequency-inverse document frequency (TF-IDF), are employed to convert text into numerical representations. Logistic Regression and Naive Bayes classifiers are then trained on this processed data.

The final model demonstrates a high level of accuracy in detecting spam news, with both Logistic Regression and Naive Bayes effectively identifying patterns indicative of spam. Naive Bayes, in particular, shows efficiency in handling high-dimensional data, making it well-suited for text classification tasks.