

# InTrAinZ Internship

## Spam News Detection and Classification

### Internship : Artificial Intelligence

Name : Y Pradeep Reddy

College : AITS Tirupati

Course : AIML

### *Topics / Agenda :-*

- Spam News Detection
- Python libraries used
- True and Fake CSV data sets
- Training and Testing Program
- Conclusion

### Spam News Detection:

#### **What is spam news detection:**

**Spam Filter** : A spam filter is a program used to detect unsolicited, unwanted and virus-infected emails and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for specific criteria on which to base its judgments.

--In spam News Detection we are training the machine first how true news looks like and how fake news looks like.

--The type of machine learning algorithm that would be used to train a system to detect spam in email messages or fake news is a supervised learning algorithm, specifically a classification algorithm.

--For Training the machine we have to take "True news" data as well as "Fake News" data in the form of ".csv" extension. which is edited in the MS Excel.  
--After training the machine it can predicts the news whether the news is true or fake.

## Python libraries used:-

### *numpy (import numpy as np):*

NumPy is a library for numerical computations in Python. It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays.

### *pandas (import pandas as pd):*

Pandas is a powerful data manipulation library. It provides data structures like DataFrame and Series, making it easy to manipulate, clean, and analyze structured data.

Statements used:

`pd.read_csv("True_news.csv")`: Reads a CSV file into a DataFrame.  
`pd.concat([dataset1, dataset2])`: Concatenates two DataFrames.

### *nltk (import nltk):*

NLTK (Natural Language Toolkit) is a library for working with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries.

Statements used:

`nltk.download('wordnet')`: Downloads the WordNet dataset

## re (import re):

The re module is Python's regular expression module, allowing for pattern matching with strings.

Statements used:

`re.sub('[^a-zA-Z]', ' ', row)`: Substitutes non-alphabetic characters with spaces.

## stopwords (from nltk.corpus import stopwords):

NLTK's stopwords module provides a list of common words (e.g., "the," "and") that are often removed from text during natural language processing tasks.

Statements used:

`stopwords.words('english')`: Retrieves a list of English stopwords.

## sklearn (from sklearn...):

Scikit-learn is a machine learning library in Python. It provides simple and efficient tools for data analysis and modeling, including various algorithms for classification, regression, clustering, and more.

Statements used:

`from sklearn.feature_extraction.text import TfidfVectorizer` TF-IDF vectorizer for converting text data into numerical vectors.

`from sklearn.model_selection import train_test_split` splits data into training and testing sets.

`from sklearn.naive_bayes import MultinomialNB` Multinomial Naive Bayes classifier.

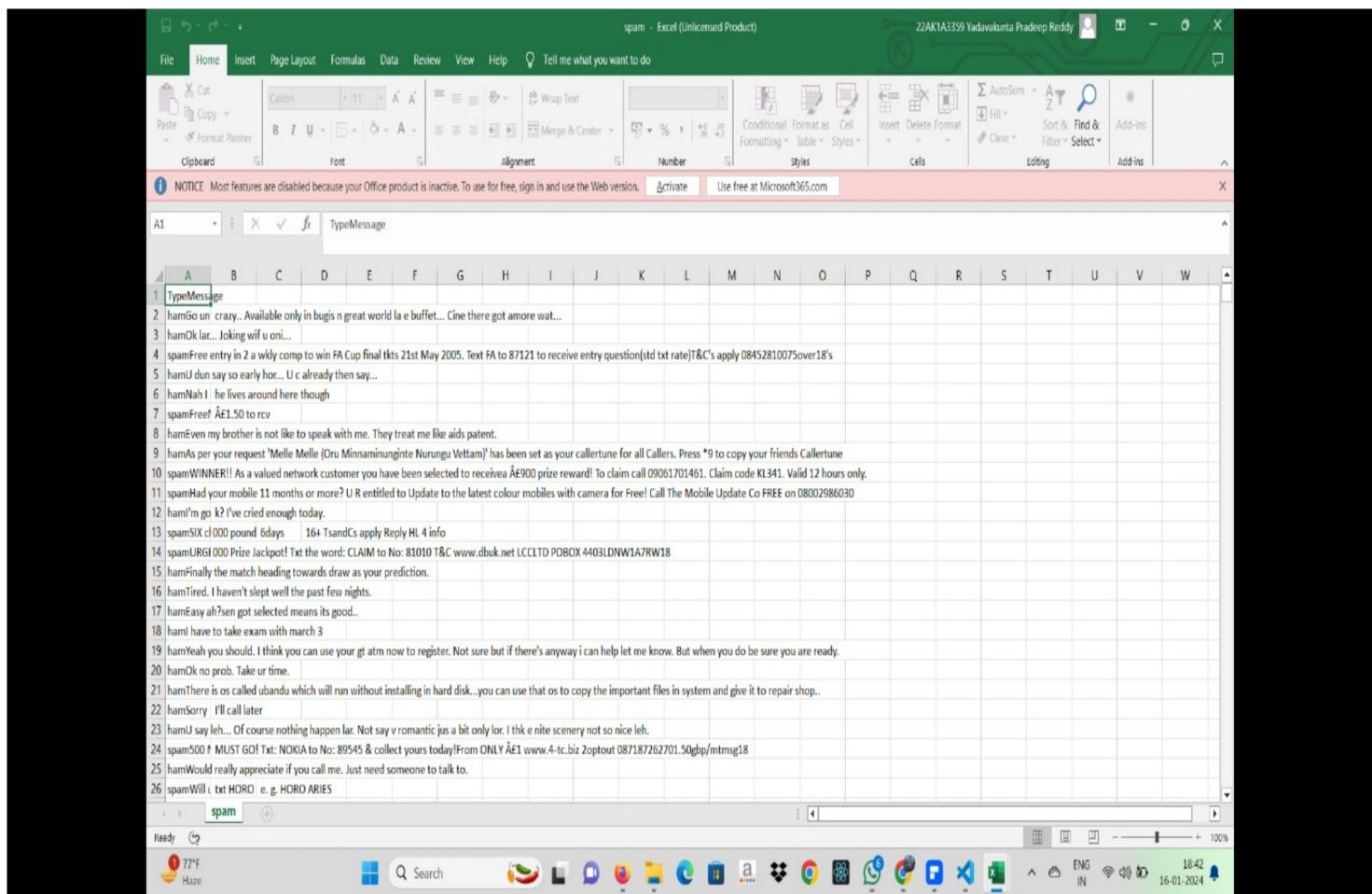
`from sklearn.metrics import accuracy_score` Computes the accuracy of the classification.

These libraries are used to perform tasks such as handling, text processing, feature extraction, machine learning model training, and performance evaluation in the provided program.

## *True and Fake CSV data sets:-*

### *For True CSV data set:*

Github link : <https://github.com/ShyamPradeepReddy/Spam-News-Detection/blob/main/True.csv>



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	TypeMessage																					
2	hamGo on crazy.. Available only in bugs n great world la e buffet... Cine there got amore wat...																					
3	hamOk lat... Joking wif u onli...																					
4	spamFree entry in 2 wkly comp to win FA Cup final tkt 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)?&C's apply 08452810075over18's																					
5	hamI dun say so early hor... U c already then say...																					
6	hamNah I he lives around here though																					
7	spamFree! £1.50 to rcv																					
8	hamEven my brother is not like to speak with me. They treat me like aids patient.																					
9	hamAs per your request 'Melli Melle (Oru Minnaminigte Nurung Vettai)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune																					
10	spamWINNER!! As a valued network customer you have been selected to receive £900 prize reward! To claim call 09061701461. Claim code K1341. Valid 12 hours only.																					
11	spamHad your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030																					
12	hamI'm go k? I've cried enough today.																					
13	spamSix cl000 pound 6days 16+ TsandCs apply Reply HL 4 info																					
14	spamURGI 000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuks.net LCL LTD POBOX 4403 LDNW1A7RW18																					
15	hamFinally the match heading towards draw as your prediction.																					
16	hamTired, I haven't slept well the past few nights.																					
17	hamEasy ah?sen got selected means its good.																					
18	hamI have to take exam with march 3																					
19	hamYeah you should, I think you can use your gt atm now to register. Not sure but if there's anyway i can help let me know. But when you do be sure you are ready.																					
20	hamOk no prob. Take ur time.																					
21	hamThere is os called ubandu which will run without installing in hard disk...you can use that os to copy the important files in system and give it to repair shop..																					
22	hamSorry I'll call later																					
23	hamI say leh... Of course nothing happen lar. Not say v romantic jus a bit only lor. I thk e nite scenery not so nice leh.																					
24	spamGO I MUST GO! Txt: NOKIA to No: 89545 & collect yours today!From ONLY £1 www.4-tc.biz 2optout 087187262701.50gbp/mmsg18																					
25	hamWould really appreciate if you call me. Just need someone to talk to.																					
26	spamWill i txt HORO e. g. HORO ARIES																					

### *For Fake CSV data set:*

Github link : <https://github.com/ShyamPradeepReddy/Spam-News-Detection/blob/main/Fake.csv>

Document Recovery

Excel has recovered the following files.  
Save the ones you wish to keep.

Fake.csv (Original)  
Version created last time the us...  
12-01-2024 07:41

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	title	text	subject	date															
2	As U.S. bu	WASHING	politicsNe	December 31, 2017															
3	U.S. milita	WASHING	politicsNe	December 29, 2017															
4	Senior U.S.	WASHING	politicsNe	December 31, 2017															
5	FBI Russia	WASHING	politicsNe	December 30, 2017															
6	Trump wa	SEATTLE	/politicsNe	December 29, 2017															
7	White Hou	WEST PALI	politicsNe	December 29, 2017															
8	Trump say	WEST PALI	politicsNe	December 29, 2017															
9	Factbox:	T	he follow	politicsNe	December 29, 2017														
10	Trump on	The follow	politicsNe	December 29, 2017															
11	Alabama c	WASHING	politicsNe	December 28, 2017															
12	Jones cert	(Reuters)	- politicsNe	December 28, 2017															
13	New York	NEW YORK	politicsNe	December 28, 2017															
14	Factbox:	T	he follow	politicsNe	December 28, 2017														
15	Trump on	The follow	politicsNe	December 28, 2017															
16	Man says	I	(in Dec. 2)	politicsNe	December 25, 2017														
17	Virginia	(Reuters)	- politicsNe	December 27, 2017															
18	U.S. laws	WASHING	politicsNe	December 27, 2017															
19	Trump on	The follow	politicsNe	December 26, 2017															
20	U.S. appes	(Reuters)	- politicsNe	December 26, 2017															
21	Treasury 5	(Reuters)	- politicsNe	December 24, 2017															
22	Federal ju	WASHING	politicsNe	December 24, 2017															
23	Exclusive:	NEW YORK	politicsNe	December 23, 2017															
24	Trump tra	(Reuters)	- politicsNe	December 23, 2017															
25	Second co	WASHING	politicsNe	December 23, 2017															
26	Failed vot	LIMA	[Re] politicsNe	December 23, 2017															

## Why CSV files are used:-

CSV (Comma-Separated Values) files are a common plain-text format used for storing tabular data. Each line in a CSV file represents a row, and the values within each row are separated by commas (or other delimiters).

## Advantages of using CSV files.

**Simplicity:** CSV files are simple and easy to understand. They are plain text, making them human-readable.

**Common Use Cases:** CSV files are widely used for data exchange between different programs, databases, and spreadsheet applications.

**Character Encoding:** CSV files are often encoded using UTF-8 or ASCII, ensuring compatibility across different systems.

*HeaderRow* :The first row often contains column names, making it easier to interpret the data.

*Quoting* Values containing special characters (like commas) are often enclosed in quotes.

*Applications* : CSV files are commonly used in data analysis, data science, and as an intermediate format for data import/export.

## Training and Testing Program:-

```
import pandas as pd
import numpy as np
True_news = pd.read_csv("True_news.csv")
#loads the True_news data into True_news
Fake_news = pd.read_csv("Fake_news.csv")
#loads the Fake_news data into Fake_news
#Add label values
True_news['label']=0
Fake_news['label']=1
dataset1 = True_news[ [ 'text' , 'label' ] ]
dataset2 = Fake_news[ [ 'text' , 'label' ] ]
#Selecting only two columns and loading
into dataset1 and dataset2
dataset = pd.concat( [ dataset1 , dataset2 ] )
#concatenating two datasets
#checking for null values
dataset.isnull().sum()                      #returns
text =0 and label 0
#For shuffling the data
dataset = dataset.sample(frac = 1)
```

```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
ps = WordNetLemmatizer()
stopwords=stopwords.words('english')
nltk.download('wordnet')

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

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

from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(max_features = 50000, lowercase=False, ngram_range=(1,2))
x=dataset.iloc[:35000,0]
y=dataset.iloc[:35000,1]

from sklearn.model_selection import train_test_split
train_data,test_data,train_label,test_label=train_test_split(x,y,test_size=0.2,random_state=0)

vec_train_data=vectorizer.fit_transform(train_data)
vec_train_data=vec_train_data.toarray()

vec_test_data=vectorizer.fit_transform(test_data)
vec_test_data=vec_test_data.toarray()
```

```
training_data=pd.DataFrame(vec_train_data,columns=vectorizer.get_feature_names()) testing_data=pd.DataFrame(vec_#Model
from sklearn.naive_bayes import MultinomialNB
clf=MultinomialNB()
clf.fit(training_data,train_label)
y_pred=clf.predict(testing_data)
from sklearn.metrics import accuracy_score
accuracy_score(test_label,y_pred)      #returns
the accuracy of prediction y_pred_train=clf.
predict(training_data)
txt="Some news we will give here"
news=clean_row(txt)
pred=clf.predict(vectorizer.transform([news]).toarray())
txt=input("Enter News:")
news=clean_row(str(txt))
pred=clf.predict(vectorizer.transform([news]).toarray())
if pred == 0:
print("News is correct")
else:
print("News is fake")
```

## Steps involved in training the above model:-

*Import the pandas and numpy libraries*

```
import pandas as pd
import numpy as np
```

*Read a CSV file named "True\_news.csv" into a DataFrame called true\_news .*

```
True_news = pd.read_csv("True_news.csv")
```

*Read a CSV file named "Fake\_news.csv" into DataFrame called Fake\_news .*

```
Fake_news = pd.read_csv("Fake_news.csv")
```

*Add labels to the news.*

```
True_news['label']=0  
Fake_news['label']=1
```

*Select only the text and label columns from both True\_news and Fake\_news DataFrames and store them in dataset1 and dataset2 respectively*

```
dataset1 = True_news[['text', 'label']]  
dataset2 = Fake_news[['text', 'label']]
```

*Concatenate dataset1 and dataset2 along the rows (axis=0) to create a new DataFrame called dataset*

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

*Check for null values in the dataset DataFrame. Returns the count of null values for each column (text and label).*

```
dataset.isnull().sum()
```

*Shuffle the rows of the dataset DataFrame.*

```
dataset = dataset.sample(frac=1)
```

*Import the Natural Language Toolkit (nltk) library, regular expression library (re), and download the WordNet dataset*

```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
ps = WordNetLemmatizer()
stopwords = stopwords.words('english')
nltk.download('wordnet')
```

*Define a function `clean_row` that performs text cleaning on a given row.*

```
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
```

*Apply the `clean_row` function to the 'text' column of the dataset `DataFrame`*

```
dataset['text'] = dataset['text'].apply(lambda x: clean_
row(x))
```

*Import `TfidfVectorizer` from `scikit-learn` and create a `TfidfVectorizer` object with specified parameters*

```
from sklearn.feature_extraction.text import
TfidfVectorizer
vectorizer = TfidfVectorizer(max_features=50000, lowercase
=False, ngram_range=(1, 2))
```

*Extract the 'text' column as `input(x)` and 'label' column as `output(y)` from the first 35,000 rows of the dataset `DataFrame`*

```
x = dataset.iloc[:35000, 0]
y = dataset.iloc[:35000, 1]
```

*Split the data into training and testing sets using the train\_test\_split function.*

```
from sklearn.model_selection import train_test_split
train_data, test_data, train_label, test_label = train_test_split(x, y, test_size=0.2, random_state=0)
```

*Apply the vectorizer to the training data and convert it to a dense array.*

```
vec_train_data = vectorizer.fit_transform(train_data)
vec_train_data = vec_train_data.toarray()
```

*Apply the vectorizer to the testing data and convert it to a dense array.*

```
vec_test_data = vectorizer.fit_transform(test_data)
vec_test_data = vec_test_data.toarray()
```

*Create DataFrames for training data and testing data with feature names obtained from the vectorizer.*

```
training_data = pd.DataFrame(vec_train_data, columns=vectorizer.get_feature_names())
testing_data = pd.DataFrame(vec_test_data, columns=vectorizer.get_feature_names())
```

*Import Multinomial Naive Bayes from scikit-learn, create a classifier object and fit it to the training data.*

```
from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB()
clf.fit(training_data, train_label)
```

*Predict the labels for the testing data using the trained classifier.*

```
y_pred = clf.predict(testing_data)
```

*Calculate and print the accuracy score for the prediction on the testing data.*

```
from sklearn.metrics import accuracy_score  
accuracy_score(test_label, y_pred)
```

*Clean a sample news text and make predictions using the trained classifier.*

```
txt = "Some news we will give here"  
news = clean_row(txt)  
pred = clf.predict(vectorizer.transform([news])).toarray()
```

*For Knowing the result.*

*Take user input of news, clean it and predict whether it is considered to be fake based on the trained classifier in the result.*

```
txt = input("Enter News:")  
news = clean_row(str(txt))  
pred = clf.predict(vectorizer.transform([news])).toarray()  
if pred == 0:  
    print("News is correct")  
else:  
    print("News is fake")
```

## Conclusion:-

Spam news detection Python program processes two CSV files containing true and fake news data. It uses Pandas for data manipulation, NLTK for text preprocessing, and scikit-learn for text vectorization and classification. The code cleans and combines the datasets, shuffles the data, applies lemmatization and removal of stopwords to the text, and utilizes TF-IDF vectorization. A Multinomial Naive Bayes classifier is trained on the preprocessed text data to distinguish between true and fake news.

The program calculates and prints the accuracy of the classification on a test set. Additionally, it allows users to input news text, cleans it, and predicts whether the news is considered correct or fake based on the trained classifier. The code is an example of a simple text classification pipeline using machine learning techniques to differentiate between true and fake news.



*Created by Pradeep  
Thank you*