Fake News Detection using NLP

Phase 5 (Complete project)

Problems

Detecting fake news using Natural Language Processing (NLP) is a challenging task due to the complexities and nuances of language, as well as the evolving tactics employed by those who create fake news. Some of the key problems faced in fake news detection using NLP include:

- ➤ Lack of labeled data
- ➤ Evolving tactics
- ➤ Misinformation vs. disinformation
- ➤ Contextual understanding
- ➤ Biases and neutrality
- ➤ Multimodal content
- ➤ Adversarial attacks
- ➤ Imbalanced datasets
- ➤ Generalization
- ➤ Ethical concerns
- ➤ Privacy concerns
- ➤ Interpretability

❖To address these problems, researchers are continually developing more sophisticated NLP models and incorporating techniques like deep learning, transfer learning, and explainability to improve fake news detection accuracy and reliability. However, it remains an ongoing and evolving area of research and development.

Dataset link:

https://www.kaggle.com/datasets/clmentbisaillon/fake-andreal-news-dataset

IMPORTANCE

- Preserving trust in information
- Preventing harm
- Minimizing the spread of false information
- Business and brand protection
- Scientific and academic integrity

DATA PREPROCESSING

- Cleaning
- Totenization
- Stop word removal
- Stemming/lemmatization
- Handling missing data

FEATURE EXTRACTION

Techniques

- ✓ **TF-IDF (Term Frequency-Inverse Document Frequency)**

- ✓ **Word Embeddings (e.g., Word2Vec)**

MODEL SELECTION

- ✓ Random forest classifier
- ✓ Logistic regression

FLOWCHART
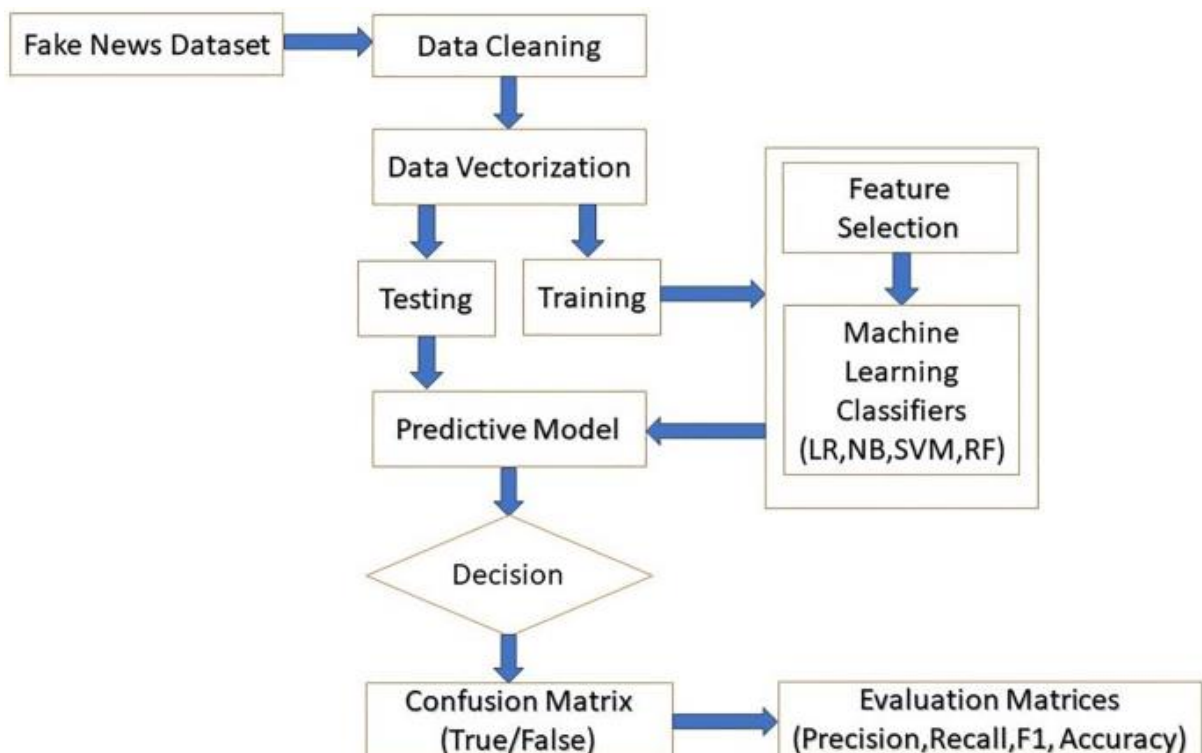
IMPORTING NECESSARY LIBRARIES


```python
import warnings
warnings.filterwarnings('ignore')
import numpy as np
import pandas as pd

import nltk
from nltk.sentiment import SentimentIntensityAnalyzer
import warnings
warnings.filterwarnings("ignore")
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
/kaggle/input/fake-and-real-news-dataset/True.csv
/kaggle/input/fake-and-real-news-dataset/Fake.csv
```

LOADING DATASET

NLP fake news detection   Draft saved

File   Edit   View   Run   Add-ons   Help

Code ▾   ● Draft Session (13h:6m)

```
[43]:   true = pd.read_csv('/kaggle/input/fake-and-real-news-dataset/True.csv')
        fake = pd.read_csv('/kaggle/input/fake-and-real-news-dataset/Fake.csv')
```

```
[44]:   fake['Category'] = 'fake'
        fake
```

[44]:

	title	text	subject	date	Category
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	fake
...	...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	fake
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	fake
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	fake
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	fake
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	fake

23481 rows × 5 columns

```
true['Category'] = 'true'
true
```

[45]:

	title	text	subject	date	Category
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	true
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	true
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	true
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	true
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	true
...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of I...	worldnews	August 22, 2017	true
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

21417 rows × 5 columns

```
data = pd.concat([fake, true], ignore_index = True)
data
```

[46]:

	title	text	subject	date	Category
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	fake
...	...	...	...	...	...
44893	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
44894	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of I...	worldnews	August 22, 2017	true
44895	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
44896	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
44897	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

44898 rows × 5 columns

2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	fake
...	...	...	...	...	...
44893	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
44894	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	true
44895	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
44896	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
44897	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

44898 rows × 5 columns

[+ Code] [+ Markdown]

```
[47]: data.shape
```

```
[47]: (44898, 5)
```

PREPROCESSING

```
[48]: data['Category'].value_counts()
```

```
[48]: Category
fake    23481
true    21417
Name: count, dtype: int64
```

```
[49]: from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
data['Category'] = encoder.fit_transform(data['Category'])
```

```
[50]: data['Category']
```

```
[50]: 0    0
1    0
2    0
3    0
```

```python
data['Category']
```

```
[50]:  0        0
       1        0
       2        0
       3        0
       4        0
              ..
       44893    1
       44894    1
       44895    1
       44896    1
       44897    1
       Name: Category, Length: 44898, dtype: int64
```

```python
vectorizer = TfidfVectorizer()
title = vectorizer.fit_transform(data['title'])
title
```

```
[51]: <44898x20896 sparse matrix of type '<class 'numpy.float64'>'
           with 546512 stored elements in Compressed Sparse Row format>
```

+ Code   + Markdown

```python
combined_df = pd.concat([fake_df_subset, true_df], axis = 0, ignore_index = True)
df = combined_df.sample(frac=1).reset_index(drop=True)
df.head(10)
```

	title	text	subject	date	label
0	Georgia Republican Ruthlessly Blocks Rape Kit...	Proving once again that Republicans don t care...	News	March 16, 2016	0
1	WOW! Former Professional Boxer Wearing "Soldie...	In 2015, former professional boxer Anthony Sma...	left-news	Apr 25, 2017	0
2	Here Are Photos Of Detroit's Public Schools T...	While much of the nation s attention has been ...	News	January 17, 2016	0
3	HORRIBLE News For Do-Nothing GOP: Americans S...	Republican lawmakers oppose paid family leave....	News	May 20, 2016	0
4	Karma's a Bi*ch: Birthers Go To Court To Chal...	They say karma s a bi*ch. It appears that at l...	News	January 14, 2016	0
5	Russia's Putin says ex-Soviet countries threat...	MOSCOW (Reuters) - Russian President Vladimir ...	worldnews	December 19, 2017	1

```python
combined_df = pd.concat([fake_df_subset, true_df], axis = 0, ignore_index = True)
df = combined_df.sample(frac=1).reset_index(drop=True)
df.head(10)
```

	title	text	subject	date	label
0	Georgia Republican Ruthlessly Blocks Rape Kit...	Proving once again that Republicans don t care...	News	March 16, 2016	0
1	WOW! Former Professional Boxer Wearing "Soldie...	In 2015, former professional boxer Anthony Sma...	left-news	Apr 25, 2017	0
2	Here Are Photos Of Detroit's Public Schools T...	While much of the nation s attention has been ...	News	January 17, 2016	0
3	HORRIBLE News For Do-Nothing GOP: Americans S...	Republican lawmakers oppose paid family leave....	News	May 20, 2016	0
4	Karma's a Bi*ch: Birthers Go To Court To Chal...	They say karma s a bi*ch. It appears that at l...	News	January 14, 2016	0
5	Russia's Putin says ex-Soviet countries threat...	MOSCOW (Reuters) - Russian President Vladimir ...	worldnews	December 19, 2017	1
6	Jewish Elders Warn Their Grandchildren: Vote ...	This election will ask us to choose one of the...	News	September 26, 2016	0
7	Zimbabwe's Mugabe 'glowed' with relief after h...	CHISHAWASHA, Zimbabwe (Reuters) - Robert Mugab...	worldnews	November 26, 2017	1
8	OKLAHOMA LAWMAKER BLASTED FOR SAYING: "Shouldn...	Oklahoma State Rep. John Bennett made a commen...	left-news	Aug 20, 2017	0
9	DEMOCRATS EAT THEIR OWN: Secret Service Protec...	It s really quite ironic that the guy who has ...	politics	May 31, 2016	0

```
[14]:  print(len(df[df.label == 0]), len(df[df.label == 1]))

       21417 21417
```

```
[17]:  X_train, X_test, y_train, y_test = train_test_split(df["text"], df["label"], test_size=0.2, random_state=42)
```

```
▷      print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

       (34267,) (8567,) (34267,) (8567,)
```

+ Code   + Markdown

MODEL BUILDING AND EXECUTION

MODEL BUILDING

** 1.Random Forest Classifier****

```
[19]:  pipeline = Pipeline([
               ('vectorizer', CountVectorizer(ngram_range=(3, 3), analyzer='word', max_features=5000)),
           ('classifier', RandomForestClassifier())
       ])
```

```
[22]:  pipeline.named_steps['vectorizer'].fit(X_train)
       pipeline.fit(X_train, y_train)
```

```
        ▸ CountVectorizer
       ┌─────────────────────────┐
       │▸ RandomForestClassifier │
       └─────────────────────────┘
```

+ Code   + Markdown

```
[23]:  y_pred = pipeline.predict(X_test)
```

```
[24]:  print(classification_report(y_test, y_pred))

                     precision    recall  f1-score   support

                  0       0.98      0.97      0.97      4341
                  1       0.97      0.98      0.97      4226

           accuracy                           0.97      8567
          macro avg       0.97      0.97      0.97      8567
       weighted avg       0.97      0.97      0.97      8567
```

2.Logistic Regression

+ Code   + Markdown

```python
[13]: from sklearn.model_selection import train_test_split
      from sklearn.preprocessing import MinMaxScaler
      from sklearn.metrics import confusion_matrix,classification_report
      from sklearn.pipeline import make_pipeline
      from sklearn.linear_model import LogisticRegression
      from sklearn.svm import SVC
```

+ Code   + Markdown

```python
[14]: X_train, X_test, y_train, y_test = train_test_split(concat_data.vector,
                                                         concat_data.label,
                                                         test_size = 0.2
```

```python
[14]: X_train, X_test, y_train, y_test = train_test_split(concat_data.vector,
                                                         concat_data.label,
                                                         test_size = 0.2,
                                                         random_state = 1,
                                                         stratify = concat_data.label)

      print(X_train.shape, y_train.shape)
      print(X_test.shape, y_test.shape)
```

```
(35918,) (35918,)
(8980,) (8980,)
```

```python
[15]: X_train = np.stack(X_train)
      X_test = np.stack(X_test)

      X_train
```

```
[15]: array([[-0.23301356,  4.5374684 , -2.3352568 , ...,  1.769452  ,
              -3.347071  ,  3.5270379 ],
             [-0.31101125,  0.70727515, -3.050925  , ..., -1.6768737 ,
              -1.6511288 ,  1.9876038 ],
             [ 0.39467925,  1.669555  ,  0.59880286, ...,  1.7896794 ,
```

```python
[15]: X_train = np.stack(X_train)
      X_test = np.stack(X_test)

      X_train
```

```
[15]: array([[-0.23301356,  4.5374684 , -2.3352568 , ...,  1.769452  ,
              -3.347071  ,  3.5270379 ],
             [-0.31101125,  0.70727515, -3.050925  , ..., -1.6768737 ,
              -1.6511288 ,  1.9876038 ],
             [ 0.39467925,  1.669555  ,  0.59880286, ...,  1.7896794 ,
              -2.3822052 ,  1.9330199 ],
             ...,
             [-3.259486  ,  2.728493  , -3.6598694 , ..., -0.78519315,
              -2.2183661 ,  0.13245803],
             [-0.92145747, -0.53599155,  1.6962891 , ..., -0.33306703,
              -0.24070585, -0.43264183],
             [-1.3668569 ,  1.0633429 , -0.39573893, ..., -0.5022214 ,
              -1.0951021 ,  0.8527349 ]], dtype=float32)
```

+ Code   + Markdown

```python
[16]: LogReg = make_pipeline(
          MinMaxScaler(),
          LogisticRegression(max_iter=1000)    ## Logistic_Regression_Classifier
```

```
[16]:   LogReg = make_pipeline(
            MinMaxScaler(),
            LogisticRegression(max_iter=1000)    ## Logistic_Regression_Classifier
        )
```

```
[17]:   def predict(model):
            model.fit(X_train,y_train)
            print(classification_report(y_test, model.predict(X_test)))
```

```
        predict(LogReg)
                      precision    recall  f1-score   support

                   0       1.00      1.00      1.00      4696
                   1       1.00      1.00      1.00      4284

            accuracy                           1.00      8980
           macro avg       1.00      1.00      1.00      8980
        weighted avg       1.00      1.00      1.00      8980
```

+ Code   + Markdown

- Thus, the dataset has been loaded, preprocessed, trained, modelled and executed successfully.

CONCLUSION

❖ Fake news has increased in recent years and it has caused a lot of harm to the society.

❖ This research project aimed to develop a model using the techniques of NLP and ML to detect if a news article/headline is fake or not and identify which methods give better output.

❖ Finding the accuracy and credibility of information and news that is available on the internet is critical nowadays.

❖ It has recently been discovered that various online platforms significantly influence disseminating misleading information and spreading fake news to serve several dreadful purposes and benefit many people.

❖ Because of the plethora of spreading and sharing data on the internet, there is a growing demand for automated false news identification systems that are accurate and efficient.

❖ A future extension of this work can be to employ attention-based deep learning approaches.