FAKE NEWS DETECTION USING NLP

TEAM MEMBER

922321106006–ARUMUGAM S

PHASE 4: DEVELOPMENT PART 2

INTRODUCTION:

In the era of information explosion, the dissemination of fake news and disinformation has become a significant challenge. The deliberate spread of misleading or false information through various media platforms can have far-reaching consequences, affecting public opinion, trust in media, and even influencing important societal decisions. As a result, the need for automated tools to detect and combat fake news has never been more critical.

Machine learning, a subset of artificial intelligence, offers a promising avenue for addressing this issue. By leveraging the power of algorithms, models, and data-driven insights, machine learning can play a vital role in fake news detection. The selection of an appropriate machine learning algorithm is a pivotal decision in this process, as it directly influences the accuracy and effectiveness of the detection system.

CONTENT FOR PHASE 4 PROJECT:

Building the project by selecting a machine learning algorithms, training the model, and evaluating its performance.

Dataset link: *https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset*

BY SELECTING MACHINE LEARNING ALGORITHM:

Detecting fake news using machine learning algorithms is a challenging but important task. Various machine learning techniques can be employed for this purpose. Here's a step-by-step guide on how to select machine learning algorithms for fake news detection:

Data Collection and Preprocessing:

- Gather a labeled dataset of news articles with labels indicating whether they are real or fake.
- Preprocess the text data by removing stop words, punctuation, and performing tokenization.

Feature Extraction:

- Convert the text data into numerical features that can be used by machine learning algorithms. Common techniques include TF-IDF (Term Frequency-Inverse Document Frequency), Word Embeddings (e.g., Word2Vec or GloVe), and N-grams.

Selecting Machine Learning Algorithms:

The choice of machine learning algorithms depends on the nature of your dataset and the complexity of the problem. You can start with the following algorithms:

- a. **Naive Bayes:**

Naive Bayes algorithms, like Multinomial Naive Bayes, are simple and work well with text data. They are efficient and can be a good baseline.

- b. **Logistic Regression:**

Logistic regression is another simple yet effective algorithm for text classification tasks.

- c. **Random Forest:**

Random Forest is an ensemble algorithm that can capture complex relationships in the data. It's robust and works well for both small and large datasets.

- d. **Support Vector Machines (SVM):**

SVM is suitable for binary classification tasks. It can work well with text data when used with appropriate kernel functions.

- e. **Neural Networks:**

Deep learning models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), can capture intricate patterns in text data but require larger datasets and computational resources.

- f. **XGBoost or LightGBM:**

These gradient boosting algorithms are excellent for classification tasks and can handle imbalanced datasets well.

- g. **Ensemble Learning:**

Combining multiple models (e.g., stacking or blending) can often lead to better results. For instance, you can combine the predictions of several algorithms to improve overall accuracy.

- **Model Evaluation:**

Split your dataset into training and testing sets (or use cross-validation) to evaluate the performance of each algorithm. Common evaluation metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

- **Hyperparameter Tuning:**

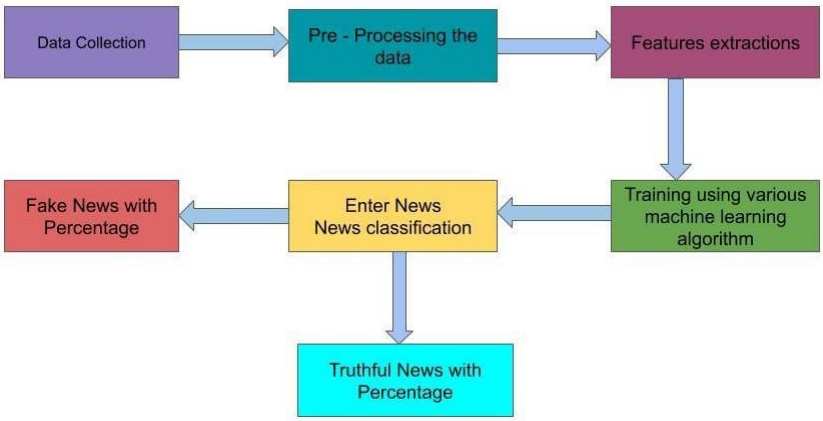
Fine-tune the hyperparameters of your selected algorithms to improve their performance. Techniques like grid search or random search can be helpful.

- **Feature Engineering:**

Experiment with different text features and their combinations to find the best representation for your problem.

- **Addressing Imbalanced Data:**

If your dataset is imbalanced (i.e., there are significantly more real news articles than fake ones), consider techniques like oversampling, undersampling, or using different evaluation metrics to handle this issue.



- **Regularization and Validation:**

Implement regularization techniques like dropout or L2 regularization in neural networks. Validate your models on unseen data to check for overfitting.

- **Post-processing:**

Depending on your model's results, you might need post-processing techniques to improve the final output.

- **Continuous Monitoring and Updating:**

Fake news evolves, so your model should be continuously updated with new data and retrained to maintain its accuracy.

Importing Libraries

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import re
import string
```

Importing Dataset

```
df_fake = pd.read_csv("../input/fake-news-detection/Fake.csv")
df_true = pd.read_csv("../input/fake-news-detection/True.csv")
```

```
df_fake.head()
```

OUTPUT:

	subject	date			
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of	politicsNews	December 31, 2017	

		a conservat...		
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	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

Inserting a column "class" as target feature

```
df_fake["class"] = 0
df_true["class"] = 1
```

```
df_fake.shape, df_true.shape
```

```
((23481, 5), (21417, 5))
```

```
# Removing last 10 rows for manual testing
df_fake_manual_testing = df_fake.tail(10)
for i in range(23480,23470,-1):
    df_fake.drop([i], axis = 0, inplace = True)

df_true_manual_testing = df_true.tail(10)
for i in range(21416,21406,-1):
    df_true.drop([i], axis = 0, inplace = True)
```

In [8]:

```
df_fake.shape, df_true.shape
```

Out[8]:

```
((23471, 5), (21407, 5))
```

```
df_fake_manual_testing["class"] = 0
df_true_manual_testing["class"] = 1
```

```
/opt/conda/lib/python3.7/site-
packages/ipykernel_launcher.py:1:
```

SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a Data Frame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

linkcode
df_fake_manual_testing.head(10)

	title	text	subject	date	class

23471	Seven Iranians freed in the prisoner swap have...	21st Century Wire says This week, the historic...	Middle-east	January 20, 2016	0
23472	#Hashtag Hell & The Fake Left	By Dady Chery and Gilbert MercierAll writers ...	Middle-east	January 19, 2016	0
23473	Astroturfing: Journalist Reveals Brainwashing ...	Vic Bishop Waking TimesOur reality is carefull...	Middle-east	January 19, 2016	0
23474	The New American Century: An Era of Fraud	Paul Craig RobertsIn the last years of the 20t...	Middle-east	January 19, 2016	0
23475	Hillary Clinton: ‘Israel First’ (and no peace ...	Robert Fantina CounterpunchAlthough the United...	Middle-east	January 18, 2016	0

23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan : US and Allied ‘Safe Zone’ Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by	21st Century Wire says As	Middle-east	January 12, 2016	0

	Iranian Military ...	21WIRE predicted in ...			

Merging True and Fake Dataframes

In [13]:

```
df_merge = pd.concat([df_fake, df_true], axis =0 )
df_merge.head(10)
```

Out[13]:

	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So	On Christmas	News	December 29, 2017	0

	Obsessed He Even Has Obama's Name...	day, Donald Trump announce d that ...			
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017	0
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017	0
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017	0
8	Former CIA Director Slams Trump	Many people have raised the alarm	News	December 22, 2017	0

	Over UN Bully...	regarding th...			
9	WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017	0

```
df_merge.columns
```

```
Index(['title', 'text', 'subject', 'date', 'class'], dtype='object')
```

Removing columns which are not required

```
df = df_merge.drop(["title", "subject","date"], axis = 1)
```

I

```
df.isnull().sum()
```

OUTPUT:

```
text     0
class    0
dtype: int64
```

Random Shuffling the dataframe

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

```
df.head()
```

OUTPUT:

	text	class
5099	During a live CNN interview with Rudy Giuliani...	0
1345	ANKARA (Reuters) - Turkey urged the United Sta...	1
20864	The attitudes of the family members defending ...	0
971	WASHINGTON (Reuters) - Charges brought against...	1
21217	The jurors in the Freddie Gray case were deadl...	0

```
df.reset_index(inplace = True)
df.drop(["index"], axis = 1, inplace = True)

df.columns

Index(['text', 'class'], dtype='object')

df.head()
```

OUTPUT:

	text	class
0	During a live CNN interview with Rudy Giuliani...	0
1	ANKARA (Reuters) - Turkey urged the United Sta...	1
2	The attitudes of the family members defending ...	0
3	WASHINGTON (Reuters) - Charges brought against...	1
4	The jurors in the Freddie Gray case were deadl...	0

Creating a function to process the texts

```
def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
df["text"] = df["text"].apply(wordopt)
```

Defining dependent and independent variables

```
x = df["text"]
y = df["class"]
```

Splitting Training and Testing

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

Convert text to vectors

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)
```

Logistic Regression

```
from sklearn.linear_model import LogisticRegression

LR = LogisticRegression()
LR.fit(xv_train,y_train)
```

```
LogisticRegression()
```

```
pred_lr=LR.predict(xv_test)
```

```
LR.score(xv_test, y_test)
```

```
0.9885026737967915
```

```
print(classification_report(y_test, pred_lr))
```

OUTPUT:

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5853
           1       0.99      0.99      0.99      5367

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

Decision Tree Classification

```
from sklearn.tree import DecisionTreeClassifier

DT = DecisionTreeClassifier()
DT.fit(xv_train, y_train)
```

```
DecisionTreeClassifier()
```

```
pred_dt = DT.predict(xv_test)
```

```
DT.score(xv_test, y_test)
```

```
0.996524064171123
```

```
print(classification_report(y_test, pred_dt))
```

OUTPUT:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5853
1	1.00	1.00	1.00	5367
accuracy			1.00	11220
macro avg	1.00	1.00	1.00	11220
weighted avg	1.00	1.00	1.00	11220

Gradient Boosting Classifier

```
from sklearn.ensemble import GradientBoostingClassifier

GBC = GradientBoostingClassifier(random_state=0)
GBC.fit(xv_train, y_train)
```

```
GradientBoostingClassifier(random_state=0)
```

```
pred_gbc = GBC.predict(xv_test)

GBC.score(xv_test, y_test)
```

```
0.9959893048128342
```

```
print(classification_report(y_test, pred_gbc))
```

	precision	recall	f1-score	support
0	1.00	0.99	1.00	5853
1	0.99	1.00	1.00	5367

accuracy			1.00	11220
macro avg	1.00	1.00	1.00	11220
weighted avg	1.00	1.00	1.00	11220

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier

RFC = RandomForestClassifier(random_state=0)
RFC.fit(xv_train, y_train)
```

```
RandomForestClassifier(random_state=0)
```

```
pred_rfc = RFC.predict(xv_test)
```

```
RFC.score(xv_test, y_test)
```

```
0.9941176470588236
```

```
print(classification_report(y_test, pred_rfc))
```

	precision	recall	f1-score	support
0	0.99	1.00	0.99	5853
1	1.00	0.99	0.99	5367

accuracy			0.99	11220
macro avg	0.99	0.99	0.99	11220
weighted avg	0.99	0.99	0.99	11220

Model Testing

```
def output_lable(n):
    if n == 0:
        return "Fake News"
    elif n == 1:
        return "Not A Fake News"

def manual_testing(news):
    testing_news = {"text":[news]}
    new_def_test = pd.DataFrame(testing_news)
    new_def_test["text"] = new_def_test["text"].apply(wordopt)
    new_x_test = new_def_test["text"]
    new_xv_test = vectorization.transform(new_x_test)
    pred_LR = LR.predict(new_xv_test)
    pred_DT = DT.predict(new_xv_test)
    pred_GBC = GBC.predict(new_xv_test)
    pred_RFC = RFC.predict(new_xv_test)

    return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGBC
Prediction: {} \nRFC Prediction:
{}".format(output_lable(pred_LR[0]),
output_lable(pred_DT[0]),

output_lable(pred_GBC[0]),
```

```
output_lable(pred_RFC[0]))))

news = str(input())
manual_testing(news)
```

BRUSSELS (Reuters) - NATO allies on Tuesday welcomed President Donald Trump s decision to commit more forces to Afghanistan, as part of a new U.S. strategy he said would require more troops and funding from America s partners. Having run for the White House last year on a pledge to withdraw swiftly from Afghanistan, Trump reversed course on Monday and promised a stepped-up military campaign against Taliban insurgents, saying: Our troops will fight to win .

U.S. officials said he had signed off on plans to send about 4,000 more U.S. troops to add to the roughly 8,400 now deployed in Afghanistan. But his speech did not define benchmarks for successfully ending the war that began with the U.S.-led invasion of Afghanistan in 2001, and which he acknowledged had required an extraordinary sacrifice of blood and treasure .

We will ask our NATO allies and global partners to support our new strategy, with additional troops and funding increases in line with our own. We are confident they will, Trump said. That comment signaled he would further increase pressure on U.S. partners who have already been jolted by his repeated demands to step up their contributions to NATO and his description of the alliance as

obsolete - even though, since taking office, he has said this is no longer the case. NATO Secretary General Jens Stoltenberg said in a statement: NATO remains fully committed to Afghanistan and I am looking forward to discussing the way ahead with (Defense) Secretary (James) Mattis and our Allies and international partners.

NATO has 12,000 troops in Afghanistan, and 15 countries have pledged more, Stoltenberg said. Britain, a leading NATO member, called the U.S. commitment very welcome . In my call with Secretary Mattis yesterday we agreed that despite the challenges, we have to stay the course in Afghanistan to help build up its fragile democracy and reduce the terrorist threat to the West, Defence Secretary Michael Fallon said. Germany, which has borne the brunt of Trump s criticism over the scale of its defense spending, also welcomed the new U.S. plan.

Our continued commitment is necessary on the path to stabilizing the country, a government spokeswoman said. In June, European allies had already pledged more troops but had not given details on numbers, waiting for the Trump administration to outline its strategy for the region.Nearly 16 years after the U.S.-led invasion - a response to the Sept. 11 attacks which were planned by al Qaeda leader Osama bin Laden from Afghanistan - the country is still struggling with weak central government and a Taliban insurgency. Trump said he shared the frustration of the American people who were weary of war without victory , but a hasty

withdrawal would create a vacuum for groups like Islamic State and al Qaeda to fill.

LR Prediction: Not A Fake News
DT Prediction: Not A Fake News
GBC Prediction: Not A Fake News
RFC Prediction: Not A Fake News

INPUT:

```
news = str(input())
manual_testing(news)
```

Vic Bishop Waking TimesOur reality is carefully constructed by powerful corporate, political and special interest sources in order to covertly sway public opinion. Blatant lies are often televised regarding terrorism, food, war, health, etc. They are fashioned to sway public opinion and condition viewers to accept what have become destructive societal norms.

The practice of manipulating and controlling public opinion with distorted media messages has become so common that there is a whole industry formed around this. The entire role of this brainwashing industry is to figure out how to spin information to journalists, similar to the lobbying of government. It is never really clear just how much truth the journalists receive because the news industry has become complacent.

The messages that it presents are shaped by corporate powers who often spend millions on advertising with the six conglomerates that own 90% of the media:General Electric (GE), News-Corp, Disney, Viacom, Time Warner, and CBS. Yet, these

corporations function under many different brands, such as FOX, ABC, CNN, Comcast, Wall Street Journal, etc, giving people the perception of choice As Tavistock s researchers showed, it was important that the victims of mass brainwashing not be aware that their environment was being controlled; there should thus be a vast number of sources for information, whose messages could be varied slightly, so as to mask the sense of external control. ~ Specialist of mass brainwashing, L. WolfeNew Brainwashing Tactic Called AstroturfWith alternative media on the rise, the propaganda machine continues to expand. Below is a video of Sharyl Attkisson, investigative reporter with CBS, during which she explains how astroturf, or fake grassroots movements, are used to spin information not only to influence journalists but to sway public opinion.

Astroturf is a perversion of grassroots. Astroturf is when political, corporate or other special interests disguise themselves and publish blogs, start facebook and twitter accounts, publish ads, letters to the editor, or simply post comments online, to try to fool you into thinking an independent or grassroots movement is speaking. ~ Sharyl Attkisson, Investigative ReporterHow do you separate fact from fiction? Sharyl Attkisson finishes her talk with some insights on how to identify signs of propaganda and astroturfing .

These methods are used to give people the impression that there is widespread support for an agenda, when, in reality, one may not exist. Astroturf tactics are also used to discredit or criticize those that disagree with certain agendas, using stereotypical names such as conspiracy theorist or quack. When in fact when someone dares to reveal the truth or questions the official story, it should spark a deeper curiosity and encourage further scrutiny of the information.This article (Journalist Reveals Tactics Brainwashing

Industry Uses to Manipulate the Public) was originally created and published by Waking Times and is published here under a Creative Commons license with attribution to Vic Bishop and WakingTimes.com. It may be re-posted freely with proper attribution, author bio, and this copyright statement. READ MORE MSM PROPAGANDA NEWS AT: 21st Century Wire MSM Watch Files

LR Prediction: Fake News
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News

INPUT:

```
news = str(input())
manual_testing(news)
```

SAO PAULO (Reuters) - Cesar Mata Pires, the owner and co-founder of Brazilian engineering conglomerate OAS SA, one of the largest companies involved in Brazil s corruption scandal, died on Tuesday. He was 68. Mata Pires died of a heart attack while taking a morning walk in an upscale district of S o Paulo, where OAS is based, a person with direct knowledge of the matter said.

Efforts to contact his family were unsuccessful. OAS declined to comment. The son of a wealthy cattle rancher in the northeastern state of Bahia, Mata Pires links to politicians were central to the expansion of OAS, which became Brazil s No. 4 builder earlier this decade, people familiar with his career told Reuters last year.

His big break came when he befriended Antonio Carlos Magalh es, a popular politician who was Bahia governor several times, and eventually married his daughter Tereza. Brazilians joked that OAS stood for Obras Arranjadas pelo Sogro - or Work Arranged by the Father-In-Law.

After years of steady growth triggered by a flurry of massive government contracts, OAS was ensnared in Operation Car Wash which unearthed an illegal contracting ring between state firms and builders. The ensuing scandal helped topple former Brazilian President Dilma Rousseff last year. Trained as an engineer, Mata Pires founded OAS with two colleagues in 1976 to do sub-contracting work for larger rival Odebrecht SA - the biggest of the builders involved in the probe.

Before the scandal, Forbes magazine estimated Mata Pires fortune at $1.6 billion. He dropped off the magazine s billionaire list in 2015, months after OAS sought bankruptcy protection after the Car Wash scandal.

While Mata Pires was never accused of wrongdoing in the investigations, creditors demanded he and his family stay away from the builder s day-to-day operations, people directly involved in the negotiations told Reuters at the time. He is survived by his wife and his two sons.

LR Prediction: Not A Fake News
DT Prediction: Not A Fake News
GBC Prediction: Not A Fake News
RFC Prediction: Not A Fake News

Conclusion:

The detection of fake news using machine learning algorithms has become an imperative undertaking in the modern information landscape. In this journey to build and train effective fake news detection models, we have explored the critical steps and considerations that form the foundation of these systems. As we conclude this discussion, it is evident that the battle against misinformation is not only vital but also dynamic, requiring adaptability, vigilance, and a commitment to ethical standards.

Through this exploration, we have discovered that the selection of the most suitable machine learning algorithm is a pivotal decision. The choice hinges on factors such as the nature of the dataset, the complexity of the problem, and available computational resources. From the simplicity of Naive Bayes to the complexity of neural networks, each algorithm offers unique advantages that can be leveraged for fake news detection.

The training process involves data collection, preprocessing, feature extraction, and meticulous model development. While the algorithms play a crucial role, the quality and diversity of the dataset, as well as hyperparameter tuning, have a profound impact on the model's effectiveness. Rigorous evaluation and testing are essential, as the real-world consequences of misclassifications can be significant.

In conclusion, the fight against fake news is an ongoing battle that necessitates collaboration among researchers, organizations, and the broader society. Machine learning algorithms, when applied judiciously, have the potential to bolster the arsenal of tools available to combat the dissemination of misinformation. However, these tools are only as effective as the diligence and ethical rigor with which they are developed and deployed.

The future of fake news detection will undoubtedly witness advancements in algorithms, data sources, and model interpretability. As such, it is paramount to remain at the forefront of innovation and research, adapting to new challenges, and ensuring that the quest for truth and accuracy in information prevails in an era awash with digital content.