IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

VOLUME : 8 ISSUE : 11 Print / Issue Publication Date: 30-Apr-2024

ISSN : 2455-2143

Indexed In

Google SCHOLAR OPEN ACCESS doi Crossref

WWW.IJEAST.COM

editor@ijeast.com

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

A SURVEY ON FAKE NEWS DETECTION USING MACHINE LEARNING ALGORITHMS

Mandeep Katre
Department of Computer Science Engineering
Inderprastha Engineering College Ghaziabad (U.P), India

Shani Yadav
Department of Computer Science Engineering
Inderprastha Engineering College Ghaziabad (U.P), India

Sumit Dwivedi
Department of Computer Science Engineering
Inderprastha Engineering College Ghaziabad (U.P), India

Abstract— **False or deceptive material that is reported as news is referred to as fake news. The enormous accessibility of digital media in recent years has made it simpler for false information to spread quickly around the world, potentially affecting public opinion and influencing political debate. The legitimacy of reliable news sources is threatened by this phenomenon, which also has the potential to erode public confidence in institutions. Researchers and policymakers are debating how to handle the fake news issue, with some calling for better media literacy instruction, more regulation of social media platforms, and increased cooperation between governments, media organizations, and civil society organizations. Despite these initiatives, the problem of fake news persists, and it will require ongoing attention and innovation.**

Keywords—**Machine learning approaches, Fake News Detection, Regression Models, Pandas, Scikit-Learn.**

I. INTRODUCTION

Creating precise and trustworthy algorithms that can recognize and categorize news articles, social media messages, and other types of online information as either fake or authentic is the challenge of fake news identification using machine learning. As disinformation and propaganda have proliferated due to how simple it is to distribute information online, the issue has grown in importance in recent years.

The purpose of fake news detection is to help people tell the difference between dependable and inaccurate information and to encourage the spread of accurate and factual information. It is necessary to apply sophisticated machine learning techniques and algorithms to solve this challenging challenge, as well as to create sizable, labelled datasets that can be used to train and test these algorithms. Accurately identifying and categorizing bogus news is one of the biggest obstacles in fake news detection. Sensationalized headlines and made-up articles are just two examples of fake news, which may be disseminated through a number of various means like social media, websites, and email. In order to identify fake news, it is necessary to utilize powerful algorithms that can analyze vast amounts of data and spot patterns and trends.

The requirement to strike a balance between accuracy and speed and efficiency in fake news detection is another difficulty. In order to stop the spread of fake news and lessen its impact, it is crucial to be able to identify it accurately as well as swiftly and effectively. This calls for the employment of algorithms that can be quickly and simply deployed across a wide range of platforms and applications, and that are accurate and scalable and provide better results.

II. METHODOLOGY

A. TF-IDF VECTORIZER

A well-liked technique in natural language processing for transforming text documents into numerical vectors that may be input into machine learning models is called TF-IDF (Term Frequency-Inverse Document Frequency). Each word in a document receives a score from the TF-IDF vectorizer that accounts for both its frequency inside the text and its significance within the corpus of documents as a whole. The document's most important features are then captured in a vector representation of the document using this score.

The term frequency (TF) of each word in a text, or simply the number of times the word appears in the document, is first determined by the TF-IDF vectorizer. However, as some words (such "the," "and," and "in") may be quite prevalent throughout all texts, the raw term frequency alone

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

may not be a fair indicator of the significance of a word in a specific document. As a result, the TF-IDF vectorizer also takes into account each word's inverse document the TF-IDF vectorizer creates a vector representation of the document that captures the relative relevance of each word in the context of the corpus by integrating the TF and IDF scores for each word. Machine learning algorithms can then utilize this vector as input to perform tasks like text categorization, clustering, or information retrieval. Because it is easy to build, effective, and generates high-quality feature representations that can raise the accuracy of machine learning models, the TF-IDF vectorizer is frequently employed in natural language processing applications.

$$idf_i = \log(\frac{n}{df_i})$$

Fig. 1. Tf-Idf Vectorizer

B. Porter Stemming

A common stemming algorithm used in natural language processing is Porter stemming, which breaks down words into their "stem," or root, or basic form. Martin Porter created the method in 1980, and it is now widely used in text analytics, search engines, and information retrieval.

The Porter stemming algorithm gradually breaks down a word to its simplest form by applying a set of rules to it. The criteria aim to eliminate typical English suffixes and ends like "-s", "-ed", "-ing", and "-ly". The following principles, for instance, would reduce the word "jumping" to the word "jump":

1. Change leaping to jump by removing the "-ing"
2. Repeat the rule for the "-ing" suffix: leap -> jump
3. Follow the "-s" suffix rule: leap -> jump

The Porter stemming algorithm is a quick and effective way to reduce the dimensionality of text data and boost the effectiveness of activities requiring natural language processing, like topic modelling, sentiment analysis, and text categorization. The production of accurate or meaningful stems by stemming algorithms is not always guaranteed, especially for words with irregular or ambiguous forms, thus it's vital to keep this in mind. As a result, the choice to utilize stemming should be made considering the task's unique requirements as well as the data's level of quality.

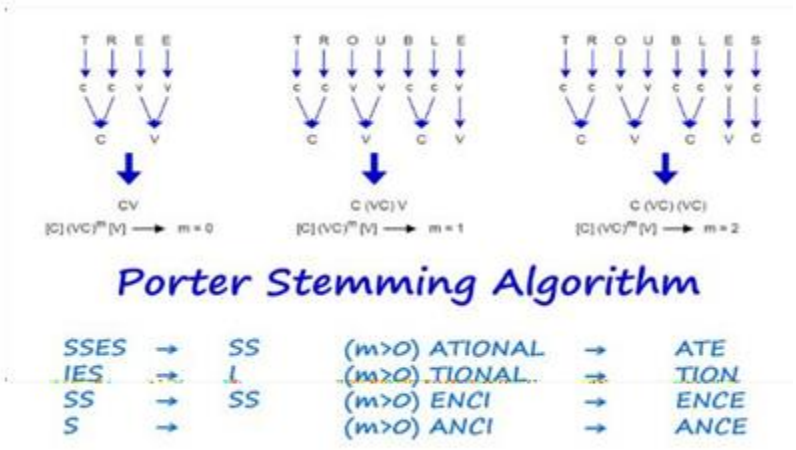

Fig.2.Porter Stemming

C. Scikit-Learn

A well-liked Python toolkit for machine learning called Scikit-learn (sometimes referred to as sklearn) offers a variety of tools and methods for data analysis and modelling. It is built upon the foundations of SciPy, NumPy, and Matplotlib and is intended to be user-friendly, effective, and extendable.

Scikit-learn includes a variety of supervised and unsupervised machine learning algorithms for classification, regression, clustering, dimensionality reduction, and model selection. It also provides tools for data preprocessing, feature extraction, and model evaluation, as well as utilities for cross-validation, pipeline construction, and ensemble methods.

Scikit-Learn functions used in this project are as follows: -

1. test_train_split () Method

Scikit-learn's train_test_split () method can be used to divide a dataset into two parts: a training set and a testing set. The technique is frequently used in supervised machine learning tasks to assess a model's performance on unobserved data.

The train_test_split () method accepts the dataset's labels and features as inputs, along with the percentage of data that will be utilized for testing (often 20% to 30%). The data is then divided at random into two subsets: one for training and one for testing. The testing set is used to assess the model's performance on fresh data while the training set is used to fit the model.

2.Accuracy_score () Method

A classification model's accuracy can be determined using the scikit-learn accuracy_score () method. It calculates the percentage of accurate predictions by comparing a model's projected labels to the test set's actual labels.

Two arrays are provided as input to the accuracy_score () method: y_true, which contains the test set's true labels, and y_pred, which contains the model's predicted labels. Then, it compares the elements of y_true and y_pred that share the same index to determine the number of accurate predictions. The accuracy score is then returned as a decimal value between 0 and 1, or as a percentage between 0% and 100%, by dividing the number of accurate predictions by the total number of forecasts.

D. Pandas

A well-known open-source Python package for data analysis and manipulation is called Pandas. It offers capabilities for data transformation, cleaning, and visualization, making it a crucial tool for analysts and data scientists.

Pandas has two main data structures: Series and DataFrame, and it is built on top of the NumPy library. A DataFrame is a two-dimensional table with rows and columns, each of which can contain a different data type. A Series is a one-dimensional array-like object that can carry data of any type.

The ability to manage missing data, its robust data aggregation and transformation skills, and its capacity to work with data from a range of sources, such as CSV and Excel files, SQL databases, and web APIs, are some of Pandas' important advantages.

Additionally, Pandas offers a variety of tools and methods for manipulating data, such as data filtering, grouping, merging, and reshaping. Additionally, it offers bar charts, scatter plots, and histograms as data visualization tools.

Overall, Pandas is a robust and adaptable data manipulation and analysis library that is widely used in data science, finance, and many other industries. It is a crucial tool for anyone using Python to interact with data because of its simple syntax and extensive capability.

E. RegEx

RegEx, or regular expressions, are strings of characters that specify a search pattern. Numerous applications, such as text processing, data validation, and search engines, employ them to match and change text data.

The re module in Python is used to manipulate regular expressions. The module offers a selection of RegEx-related functions and techniques, such as search (), findall (), and sub (). These operations can be used to extract data from strings, look for patterns in strings, and substitute new values for specific sections of a string.

The literal characters and extra characters known as metacharacters combine to form regular expressions. For instance, the asterisk (*) metacharacter matches zero or more repetitions of the preceding character or group, while the dot (.) metacharacter matches any single character.

Numerous text-manipulation tasks, including finding email addresses, checking phone numbers, and extracting information from complicated texts, can be completed using regular expressions. It can be difficult to create sophisticated regular expressions, therefore it's crucial to thoroughly test them to make sure they function as intended.

F. Logistic Regression

A statistical method for classification problems where the output variable is categorical is called logistic regression. In binary classification situations, where the output variable can have one of two potential values, such as yes or no, true or false, or 0 or 1, it is frequently employed in machine learning.

In logistic regression, the input data is first processed using a linear function, and then the output values are passed through a sigmoid function to map them to a probability value between 0 and 1. The probability that the supplied data belongs to a specific class can be calculated using this probability value.

In order for the projected probabilities of the model to match the actual probabilities of the training data, the logistic regression algorithm iteratively updates the weights and biases of the linear function using a gradient descent technique. The algorithm's goal
is to minimize the cross-entropy loss function, which gauges the discrepancy between the training data's actual labels and anticipated probabilities.

Numerous benefits of logistic regression include its simplicity, interpretability, and simplicity in application. Additionally, it can handle relationships that are both linear and nonlinear between the input variables and the output variable. However, other machine learning techniques like decision trees or support vector machines may be better effective for solving more complex categorization issues.

G. NumPy

Python's NumPy library is used to perform numerical computations. It offers strong facilities for manipulating sizable, multidimensional arrays and matrices as well as a vast array of mathematical operations for carrying out intricate calculations.

Because it is developed on top of the C programming language, NumPy is quick and effective at handling

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

numerical calculations. The ndarray, or N-dimensional array, which may represent arrays of any dimensionality, serves as its main data structure.

For working with ndarrays, NumPy offers a variety of functions and techniques, such as indexing, slicing, and resizing arrays as well as performing mathematical operations like addition, subtraction, multiplication, and division. Additionally, it offers resources for carrying out operations in linear algebra such matrix multiplication, inversion, and eigenvalue decomposition.

NumPy's ability to execute vectorized operations, which enables it to conduct calculations on large arrays at once rather than looping over individual elements, is one of its main advantages. This makes using NumPy for numerical computations significantly faster and more effective than utilizing Python's built-in data structures.

In conclusion, NumPy is a robust and adaptable Python library for numerical computing that is widely used in a variety of fields, such as scientific computing, data analysis, and machine learning. Anyone working with numerical data in Python needs to have this tool because of its simple syntax and extensive capability.

III. RESULTS AND FINDINGS

```
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
```

```
data_fake = pd.read_csv('/content/Fake.csv')
data_true = pd.read_csv('/content/True.csv')
data_fake.head()
data_true.head()
```

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

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

```
data_fake["class"] = 0
data_true['class'] = 1
```

```
data_fake.shape, data_true.shape
```

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

```
data_fake_manual_testing = data_fake.tail (10)
for i in range (23480,23470,-1):
  data_fake.drop([i], axis= 0, inplace=True)

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

```
data_fake_manual_testing['class'] - 0
data_true_manual_testing['class'] - 1
```

```
data_merge - pd.concat([data_fake, data_true], axis - 0)
data_merge.head (10)
```

	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 Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
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 Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017	0
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

```
data_merge.columns
```

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

```
data = data_merge.drop(['title', 'subject', 'date'], axis= 1)
```

```
data.isnull().sum()
```

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

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

```
data.head()
```

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

	text	class
9848	HOLLYWOOD, Fla. (Reuters) - U.S. Republican of...	1
8842	Urbandale High School students worked hard to ...	0
410	Equifax, the oldest and most valuable of the t...	0
10006	SACRAMENTO, Calif. (Reuters) - California will...	1
3606	It s funny how Trump kept screaming that the e...	0

```
data.reset_index(inplace = True)
data.drop(['index'], axis = 1, inplace = True)
```

```
data.columns
```

```
Index(['level_0', 'text', 'class'], dtype='object')
```

```
data.head()
```

	level_0	text	class
0	0	HOLLYWOOD, Fla. (Reuters) - U.S. Republican of...	1
1	1	Urbandale High School students worked hard to ...	0
2	2	Equifax, the oldest and most valuable of the t...	0
3	3	SACRAMENTO, Calif. (Reuters) - California will...	1
4	4	It s funny how Trump kept screaming that the e...	0

```
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
```

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

```
data['text'] = data['text'].apply(wordopt)
```

```
x = data['text']
y = data['class']
```

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

```
from sklearn.feature_extraction.text import TfidfVectorizer

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

```
from sklearn.linear_model import LogisticRegression

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

```
LogisticRegression()
```

```
0.9878787878787879
```

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

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5826
1	0.99	0.99	0.99	5394
accuracy			0.99	11220
macro avg	0.99	0.99	0.99	11220
weighted avg	0.99	0.99	0.99	11220

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

```
from sklearn.tree import DecisionTreeClassifier

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

```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

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

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

```
0.9954545454545455
```

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

```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      5826
           1       1.00      0.99      1.00      5394

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

```
from sklearn.ensemble import GradientBoostingClassifier

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

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

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

```
pred_gb = GB.predict(xv_test)
```

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

```
0.9951871657754011
```

```
print(classification_report(y_test, pred_gb))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      5826
           1       0.99      1.00      1.00      5394

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

```
from sklearn.ensemble import RandomForestClassifier

RF = RandomForestClassifier()
RF.fit(xv_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier()
```

```
pred_rf = RF.predict(xv_test)
```

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

```
0.9883244206773618
```

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 11, ISSN No. 2455-2143, Pages 222-231
Published Online March 2024 in IJEAST (http://www.ijeast.com)

```
print(classification_report(y_test, pred_rf))
```

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5826
1	0.99	0.98	0.99	5394
accuracy			0.99	11220
macro avg	0.99	0.99	0.99	11220
weighted avg	0.99	0.99	0.99	11220

```
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_GB = GB.predict(new_xv_test)
    pred_RF = RF.predict(new_xv_test)

    return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGB Prediction: {} \nRF Prediction: {} " .format (output_lable(pred_LR[0]),
                                                                                                              output_lable(pred_DT[0]),
                                                                                                              output_lable(pred_GB[0]),
                                                                                                              output_lable(pred_RF[0])))
```

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

```
Chief Minister Arvind Kejriwal on Sunday met former Delhi Health Minister Satyendar Jain in the hospital, describing him as a hero.  Kejriwal shared three differe


LR Prediction: Fake News
DT Prediction: Fake News
GB Prediction: Fake News
RF Prediction: Not A Fake News
```

IV. REFERENCES

[1]. Shu,K.,Mahudeswaran,D.(2017).,Wang,S.,Lee,K., &Liu, H. Data mining viewpoint on the detection of fake news on social media. 19(1), 22–36, ACM SIGKDD Explorations Newsletter.

[2]. Y.,Seo,S.,&Ruchansky,N.(2017).CSI:Ahybriddeep model for the identification of fake news. Preprint available at arXiv:1703.06959.

[3]. Zarrabi-Zadeh,H.,M.S.M.Bafghi,&F.Karimi(2019).Social mediafakenewsdetectionusingmachinelearningalgorithms.522- 534 in Expert Systems with Applications, 116.

[4]. Khan, S., Shah, S. A., and Z. A. Khan (2019). A thorough analysisoffalsenewsdetectionmethods,includingataxonomyof datasets, methods, and future research. 78(17), 23917–23963 Multimedia Tools and Applications.

[5]. Liu,J.,Wu,H.,&Co.(2018).Asurveyaboutspottingrumours via social media. TIST, 9(3), 1–38. ACM Transactions on Intelligent Systems and Technology.

[6]. Pérez-Rosas, B. Kleinberg, A. Lefevre, & R. Mihalcea (2018). detection of bogus news automatically. 28th International Conference on Computational Linguistics Proceedings, 3391-3401.

[7]. W. Y. Wang (2017). "Liar, liar, trousers on fire": A fresh benchmark dataset for spotting false news. arXiv preprint 1705.00648 is available.

[8]. Zhou, X., Zafarani(2020),R., & Shu, K. Fake news: A review of potential, detection techniques, and studies. Preprintforar Xivis arXiv:2004.12831.

IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

ABOUT IJEAST

International Journal of Engineering Applied Science and Technology (IJEAST) is a peer-reviewed, open access journal that publishes high-quality research papers in the field of Engineering, Applied Science and Technology.

IJEAST aims to provide a platform for researchers, academicians, and professionals to share their innovative ideas, research findings, and practical experiences with the global scientific community.

FOCUS AREAS

- Engineering
- Applied Science
- Technology
- Innovation & Development
- Interdisciplinary Studies

PEER REVIEWED
All submissions are rigorously peer reviewed to ensure quality.

OPEN ACCESS
Free and unrestricted access to research for all.

GLOBAL REACH
Connecting researchers and professionals worldwide.

TIMELY PUBLICATION
We ensure a swift and efficient publication process.

For more information, visit our website
www.ijeast.com

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

editor@ijeast.com
www.ijeast.com
India

2455-2143