PANIMALAR
ENGINEERING COLLEGE
An Autonomous Institution, Affiliated to Anna University, Chennai
A Christian Minority Institution
(JAISAKTHI EDUCATIONAL TRUST)
Approved by All India Council for Technical Education

Department of Computer Science And Engineering

FAKE NEWS DETECTION USING MACHINE LEARNING

TEAM MEMBER

U.THIYANESHWAR (211423104701)

Domain: Machine Learning

27-10-2025

Guide Name

Mr.VEERAMANIKANDAN K

Coordinator

Mr.SASIKUMAR A N

ABSTRACT

- **Objective**: To create an intelligent system using NLP and machine learning to accurately distinguish between real and fraudulent news postings.
- **Dataset**: The system was trained and evaluated using the public Employment Scam Aegean Dataset (EMSCAD).
- **Methods**: It uses Bag-of-Words (BoW) and TF-IDF to convert news text into numerical features for analysis.
- **Models**: The project compares four ML models: Support Vector Machine (SVM), Random Forest, Naive Bayes, and XGBoost.
- **Results:** A final ensemble model, combining SVM, RF, and NB, achieved a superior accuracy of 98.6% in detecting fake news.

OBJECTIVES

•The **primary objective** is to develop an effective recruitment fraud detection model utilizing state-of-the-art machine learning techniques.

•To develop an effective model capable of accurately distinguishing between legitimate and fraudulent news postings.

•To leverage Natural Language Processing (NLP) and supervised machine learning techniques, as traditional manual fact-checking is time-consuming and inefficient at scale.

•To employ two powerful feature extraction techniques, **Bag-of-Words (BoW)** and **Term Frequency-Inverse Document Frequency (TF-IDF)**, to convert unstructured article text into a numerical format.

•To implement, train, and compare four distinct machine learning models: **Support Vector Machine (SVM)**, **Random Forest**, **Naive Bayes**, and **XGBoost** as a state-of-the-art benchmark.

•To create an **ensemble model** that combines the predictions from the three primary classifiers (SVM, RF, and NB) using a simple majority vote to enhance predictive accuracy and robustness.

What Our Project Offers

- Intelligent Fraud Detection
- High-Accuracy Ensemble Model
- Advanced Machine Learning (SVM, Random Forest, Naive Bayes & XGBoost)
- NLP Text Analysis (TF-IDF & Bag-of-Words)
- State-of-the-Art Benchmarking (using XGBoost)
- Automated & Fast Classification
- Protects Users from Misinformation
- Reliable Solution (98.6% Accuracy)

Sustainable Development Goal (SDG) Target for This Project:

SDG 4 & 8 (Quality Education & Decent Work): The project helps create a safer environment for news seekers. It protects applicants from recruitment fraud and scams that aim to steal personal information or cause economic hardship.

SDG 3 (Good Health and Well-being): The report notes that your project's methodology (using SVM, Random Forest, etc.) could be adapted to identify and filter malicious fake news related to health, preventing the spread of misinformation .

SDG 5 (Gender Equality): By making the online hiring process safer, the system inherently protects all news applicants from fraud, regardless of gender.

SDG 9 (Industry, Innovation, and Infrastructure): Your project is an example of technological innovation applied to combat modern cybercrime. It uses advanced ML and ensemble models to build an "effective recruitment fraud detection model," which strengthens the security of online industry infrastructure.

SDG 11 (Sustainable Cities and Communities): By addressing online recruitment scams, the system helps protect the economic well-being and privacy of individuals within communities.

Problem Statement (WHY?)

The Problem: The widespread use of the World Wide Web and social media for online recruitment has created an environment ripe for fraudulent activities.

Why? Scammers exploit these platforms by posting fake news advertisements to deceive applicants, leading to consequences like:

- Theft of private information
- Financial loss
- Damage to corporate reputations

The Need: The primary problem is to develop an **automated and reliable system** that can accurately distinguish between legitimate and fraudulent news postings from a large volume of online data.

Key Challenges:

- **Handling Unstructured and Mixed Data:** News postings contain a combination of structured and unstructured free-text data (e.g., company profile, description).
- **Class Imbalance:** Datasets are inherently imbalanced, with a very small percentage of fraudulent ads compared to legitimate ones, making it hard for models to learn.
- **Feature Identification:** It is crucial to determine the most indicative features that differentiate fake posts from real ones.

Proposed Solution

- The proposed system is an effective recruitment fraud detection model designed to determine whether a news posting is genuine or fraudulent. It follows a structured workflow: Data Preprocessing, Textual Analysis, Feature Construction, Classification, and Performance Evaluation. The system extracts features from the text data using two primary techniques: Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).

- It trains and evaluates four distinct supervised machine learning models:

- Support Vector Machine (SVM)

- Random Forest

- Naive Bayes Classifier

- XGBoost (Extreme Gradient Boosting), which is included as a state-of-the-art benchmark for comparison.

- Finally, an ensemble model is created by combining the predictions from the three individual models (SVM, Random Forest, and Naive Bayes).

Implementation Plan (HOW?)

1**. Data Preprocessing**:

Load and label the EMSCAD dataset (Real=0, Fake=1) .

Clean all text: Convert to lowercase, remove stopwords, and remove punctuation/numbers .

2**. Feature Engineering:**

Convert cleaned text into numerical vectors using TF-IDF Vectorizer .

Split the data into 80% Training and 20% Testing sets .

3. **Model Training:**

Train four individual models: Random Forest, SVM, Naive Bayes, and XGBoost .

Create and train a final Ensemble Model (a VotingClassifier) that combines the predictions of RF, SVM, and NB .

4. **Evaluation**:

Test all trained models on the unseen 20% test data .

Compare performance using Accuracy, Precision, Recall, and F1-Score .

5. **Implementation**:

Save the trained TF-IDF vectorizer and the best-performing models (Ensemble, XGBoost) using joblib .

Create a final predict.py script that loads these files to classify new, raw text input .

TECHNOLOGIES USED

- **Core Language**
- Python

- **Machine Learning & Data Science Libraries**
- Scikit-learn (sklearn): Used for ML models, vectorizers, and metrics.
- Pandas: Used for data loading, manipulation, and analysis (e.g., DataFrames).
- XGBoost: Used for the high-performance XGBClassifier model.
- Joblib: Used for saving and loading the trained models .

- **Natural Language Processing (NLP)**
- NLTK (Natural Language Toolkit): Used for text processing, specifically for removing "stopwords".
- TF-IDF (Term Frequency-Inverse Document Frequency): The main technique used to convert text into numerical features.
- Bag-of-Words (BoW) / CountVectorizer: The alternative feature extraction technique mentioned.

- **Machine Learning Models & Algorithms**
- Random Forest
- Support Vector Machine (SVM)
- Naive Bayes (MultinomialNB)
- Ensemble Model (Voting Classifier): Combines the predictions of RF, SVM, and NB for a final vote.

- **Application & Deployment (As described in your report)**
- Docker: Used to package the application, models, and dependencies into a container.
- Kubernetes: Mentioned as the platform to manage and scale the Docker containers in a production environment .

System Workflow

Input Data: The process begins with the **News Posting Dataset**, which includes both genuine and fraudulent news advertisements.

Data Preprocessing: The system receives the raw posting data. It is then cleaned, formatted, and missing values are handled to create a tidy dataset suitable for analysis.

Feature Extraction and Construction: Following preprocessing, textual analysis is performed to extract relevant features from the news ads. This is done using two primary techniques: **Bag-of-Words (BoW)** and **Term Frequency-Inverse Document Frequency (TF-IDF)**.

Model Classification: The extracted features are fed into three distinct machine learning models for classification: Support Vector Machine (SVM), Random Forest, and Naive Bayes. Each model is trained to independently predict whether a news posting is fraudulent.

Ensemble Prediction: The individual predictions from the three models are then combined in an ensemble model. This model uses a simple majority vote to determine the final classification.

Output: Finally, the system outputs a prediction, classifying the news posting as either genuine or false. The model's performance is then evaluated using metrics like accuracy, precision, and recall.

Case Study

Case Study: Fake News Detection
Project: Fake News Detection using Machine Learning.
Problem: The rapid proliferation of digital media has led to a significant challenge in managing misinformation, particularly fraudulent news postings that threaten user privacy and corporate reputations. Traditional manual fact-checking is too time-consuming and inefficient to handle this large scale.

Dataset: The system was trained and evaluated on the public **Employment Scam Aegean Dataset (EMSCAD)**, which contains 17,800 news advertisements.

Methodology: The project leverages Natural Language Processing (NLP) and supervised machine learning techniques. It uses **Bag-of-Words (BoW)** and **Term Frequency-Inverse Document Frequency (TF-IDF)** to convert unstructured text into numerical features suitable for classification.

Models Implemented: Four distinct machine learning models were trained and compared**:**

- Support Vector Machine (SVM)
- Random Forest
- Naive Bayes
- XGBoost (as a state-of-the-art benchmark)

Proposed Solution: An Ensemble Model was created. This model combines the predictions from the three primary classifiers (SVM, Random Forest, and Naive Bayes) and uses a simple majority vote to make a final, more robust prediction.

Key Result: The system demonstrated high efficiency. The final Ensemble Model achieved a superior accuracy of 98.6% , proving this methodology is a reliable and modern solution for combating online misinformation**.**

Target Users & Impact

Target Users

- **Applicants / News Seekers:** Individuals looking for news, including recent graduates, who are the primary targets of fraudulent postings .
- **Companies & Corporations:** Businesses whose reputations are threatened by fraudulent news postings made in their name.
- **Individuals & Communities:** The general public whose privacy and economic well-being are at risk from online scams.

Impact

- **Combats Misinformation:** Provides a reliable and modern solution to fight the spread of online misinformation.
- **Protects Users:** Creates a safer environment for news seekers and protects applicants from recruitment fraud, data theft, and financial loss.
- **Secures Industry:** Strengthens the security of online industry infrastructure by identifying and flagging malicious content.
- **Safeguards Reputations:** Protects corporate reputations from being damaged by fraudulent news.
- **Protects Individuals:** Helps protect the economic well-being and privacy of individuals within communities.
- **High Accuracy:** The system is highly effective, with the final ensemble model achieving **98.6% accuracy**.

Future Scope / Vision

- Future iterations of the algorithm could deliver superior outcomes using hybrid ways to achieve the same goals.
- In the future, the prototype's effectiveness and accuracy can be improved to a certain extent.
- The user interface can also be improved.
- Furthermore, hybridized optimization techniques can be used to increase the model's prediction ability for detecting e-recruitment fraud

Base paper/References

Base Paper

Santhiya, P., Kavitha, S., Aravindh, T., et al. (2023). "FAKE NEWS DETECTION USING MACHINE LEARNING."

In *Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI)* .

Your report identifies this as the base paper for your project, using the EMSCAD dataset and the ensemble methodology.

Key References

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). "Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset."

This is the paper that introduced the **EMSCAD dataset** used in your project.

Ahmad, I., Yousaf, M., et al. (2020). "Fake News Detection Using Machine Learning Ensemble Methods."

Supports your use of ensemble methods for fake news detection .

Choudhary, A., & Arora, A. (2021). "Linguistic feature based learning model for fake news detection and classification."

A key paper on using linguistic features, relevant to your NLP approach .

Sharma, U., Saran, S., & Patil, S. M. (2021). "Fake News Detection using Machine Learning Algorithms."

Discusses the use of ML algorithms for this problem, similar to your own work .

PROJECT IMPROVEMENTS

- **Drawbacks / Limitations**

Class Imbalance: The dataset is naturally imbalanced, with a very small percentage of fraudulent ads compared to legitimate ones, which is a key technical challenge for training .

Subtle Feature Identification: It can be difficult for the model to identify the *subtle patterns* in textual content and organizational details that signal a scam .

Scope Limitation: The system's effectiveness is based on its training data; the report notes that "not all phoney news will spread via online networking sites".

- **Future Scope (Improvements)**

Improve Accuracy: The prototype's effectiveness and accuracy can be improved further.

Enhance User Interface: The user interface can be improved for a better user experience.

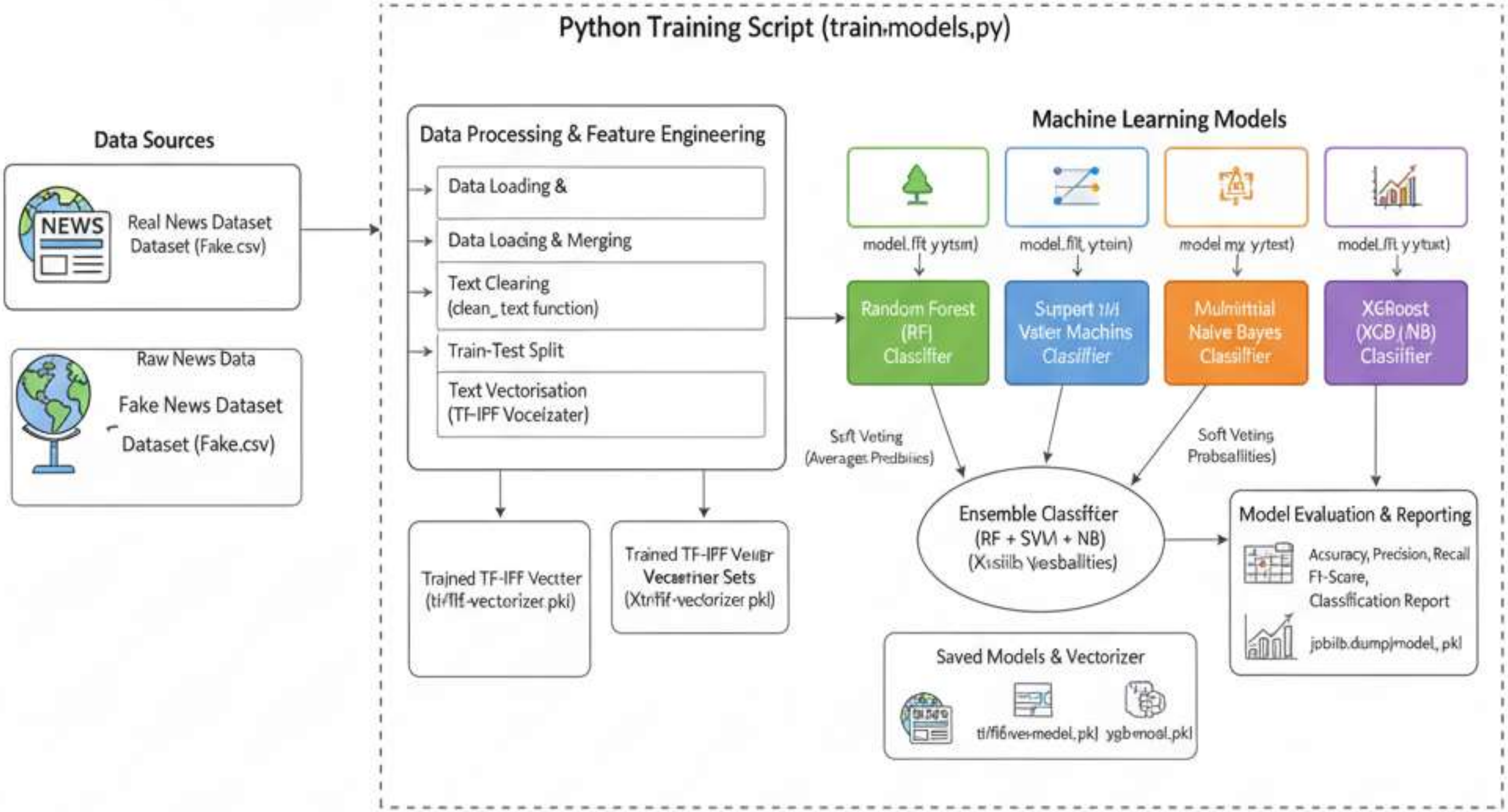
Use Hybrid Methods: Future iterations could use "hybrid ways" or "hybridized optimization techniques" to deliver superior outcomes and increase prediction ability.

BASE PAPER DETAILS

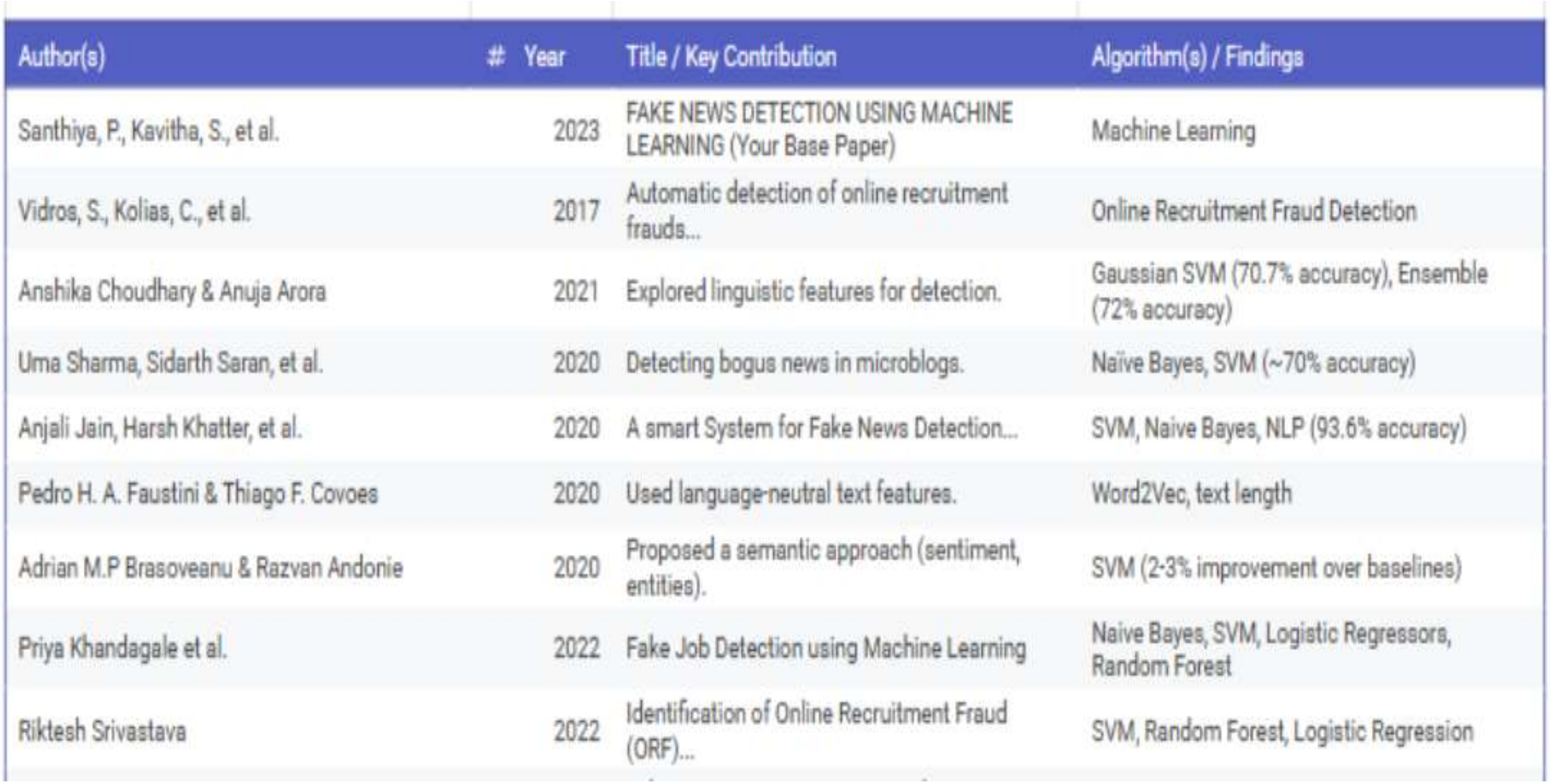
Title: "FAKE NEWS DETECTION USING MACHINE LEARNING"

Authors: Santhiya, P., Kavitha, S., Aravindh, T., Archana, S., & Praveen, A. V.
Publisher / Conference: Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI)
Link (DOI): 10.1109/ICCCI56745.2023.10128339

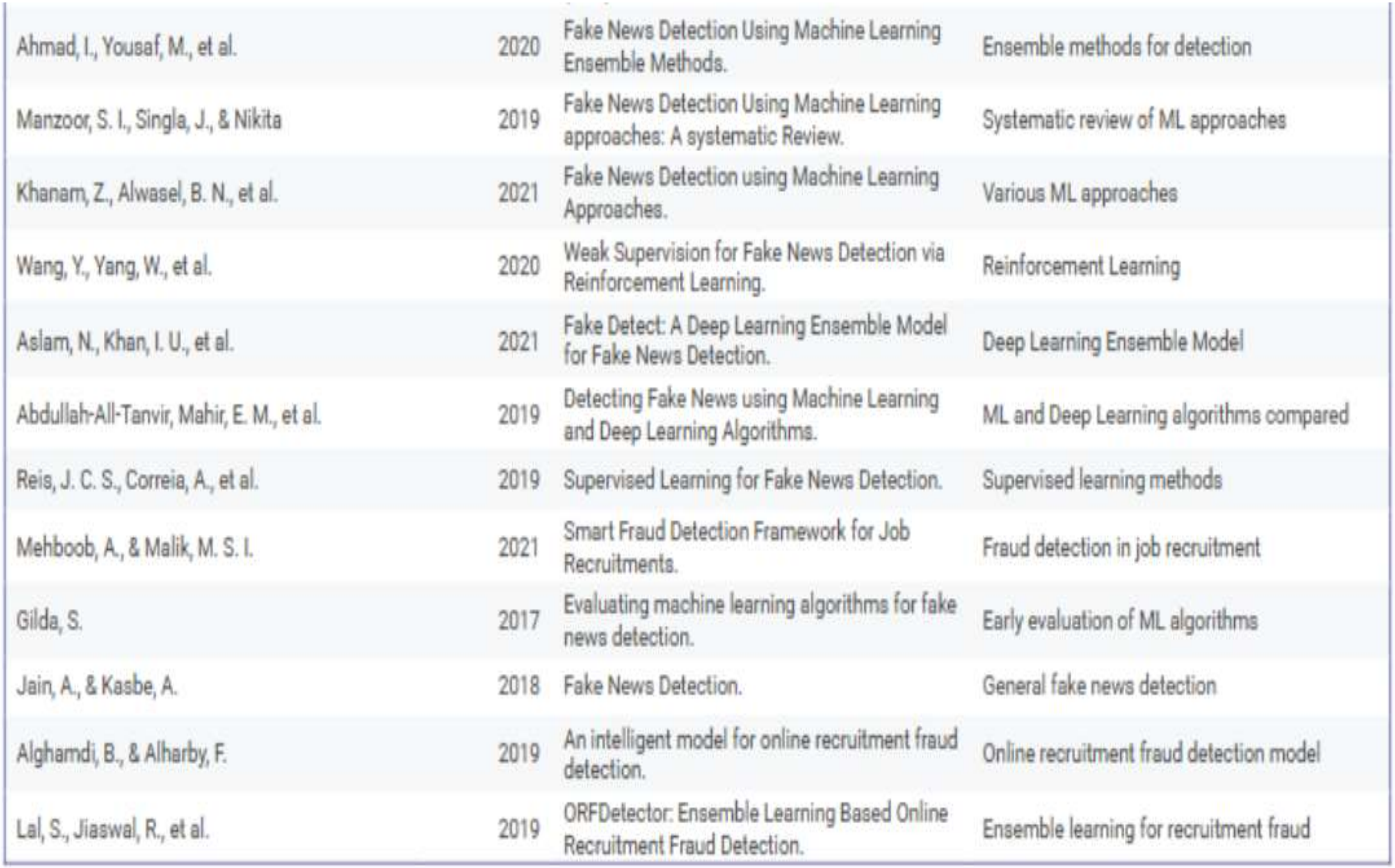
ARCHITECTURE DIAGRAM



LITERATURE SURVEY



Author(s)	#	Year	Title / Key Contribution	Algorithm(s) / Findings
Santhiya, P., Kavitha, S., et al.		2023	FAKE NEWS DETECTION USING MACHINE LEARNING (Your Base Paper)	Machine Learning
Vidros, S., Kolias, C., et al.		2017	Automatic detection of online recruitment frauds...	Online Recruitment Fraud Detection
Anshika Choudhary & Anuja Arora		2021	Explored linguistic features for detection.	Gaussian SVM (70.7% accuracy), Ensemble (72% accuracy)
Uma Sharma, Sidarth Saran, et al.		2020	Detecting bogus news in microblogs.	Naïve Bayes, SVM (~70% accuracy)
Anjali Jain, Harsh Khatter, et al.		2020	A smart System for Fake News Detection...	SVM, Naive Bayes, NLP (93.6% accuracy)
Pedro H. A. Faustini & Thiago F. Covoes		2020	Used language-neutral text features.	Word2Vec, text length
Adrian M.P Brasoveanu & Razvan Andonie		2020	Proposed a semantic approach (sentiment, entities).	SVM (2-3% improvement over baselines)
Priya Khandagale et al.		2022	Fake Job Detection using Machine Learning	Naive Bayes, SVM, Logistic Regressors, Random Forest
Riktesh Srivastava		2022	Identification of Online Recruitment Fraud (ORF)...	SVM, Random Forest, Logistic Regression



Ahmad, I., Yousaf, M., et al.	2020	Fake News Detection Using Machine Learning Ensemble Methods.	Ensemble methods for detection
Manzoor, S. I., Singla, J., & Nikita	2019	Fake News Detection Using Machine Learning approaches: A systematic Review.	Systematic review of ML approaches
Khanam, Z., Alwasel, B. N., et al.	2021	Fake News Detection using Machine Learning Approaches.	Various ML approaches
Wang, Y., Yang, W., et al.	2020	Weak Supervision for Fake News Detection via Reinforcement Learning.	Reinforcement Learning
Aslam, N., Khan, I. U., et al.	2021	Fake Detect: A Deep Learning Ensemble Model for Fake News Detection.	Deep Learning Ensemble Model
Abdullah-All-Tanvir, Mahir, E. M., et al.	2019	Detecting Fake News using Machine Learning and Deep Learning Algorithms.	ML and Deep Learning algorithms compared
Reis, J. C. S., Correia, A., et al.	2019	Supervised Learning for Fake News Detection.	Supervised learning methods
Mehboob, A., & Malik, M. S. I.	2021	Smart Fraud Detection Framework for Job Recruitments.	Fraud detection in job recruitment
Gilda, S.	2017	Evaluating machine learning algorithms for fake news detection.	Early evaluation of ML algorithms
Jain, A., & Kasbe, A.	2018	Fake News Detection.	General fake news detection
Alghamdi, B., & Alharby, F.	2019	An intelligent model for online recruitment fraud detection.	Online recruitment fraud detection model
Lal, S., Jiaswal, R., et al.	2019	ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection.	Ensemble learning for recruitment fraud

MODULES

Module 1: Data Preprocessing and Feature Engineering

This module transforms the raw news posting text into a clean, numerical format that the machine learning models can understand.

Key Points:

Data Input: Uses the **EMSCAD dataset** (17,800 news ads).

Data Cleaning: Cleans the text by converting it to lowercase, removing punctuation, and filtering out common "stopwords" (like 'the', 'is', 'a') .

Feature Construction: New features are created, such as text length or whether a company profile is missing.

Examples of Techniques:

Bag-of-Words (BoW): Represents text as a collection of its word frequencies, ignoring grammar.

Term Frequency-Inverse Document Frequency (TF-IDF): A statistical metric that evaluates how important a word is to a specific document in relation to the entire collection of documents.

Module 2: Classification Models

This module contains the core predictive logic. It uses several machine learning algorithms to classify the news postings as either "Genuine" or "Fraudulent" .

Key Points:

It implements, trains, and compares different supervised learning models.

It includes a high-performance model as a benchmark.

Examples of Models Used:

Random Forest: An ensemble model that builds multiple decision trees and uses their collective vote. It's effective on large datasets and achieved over **98.18% accuracy** in your project.

Support Vector Machine (SVM): A classifier that finds the optimal boundary (hyperplane) to separate data points into two distinct classes .

Naive Bayes: A classifier based on Bayes' Theorem, which is highly effective for text classification tasks .

XGBoost: An advanced and highly accurate gradient boosting model used as a **state-of-the-art benchmark** to compare the other models against.

Module 3: Ensemble and Evaluation

This final module integrates the predictions from the individual models to create a more robust final answer and evaluates the system's overall performance .

Key Points:

It combines the strengths of multiple models to improve accuracy .

It uses standard metrics to measure how effective the models are.

Examples:

Ensemble Model: The system combines the predictions from **SVM, Random Forest, and Naive Bayes**. It uses a **simple majority vote** to make the final decision. For example, if two out of three models classify a post as "Fraudulent," the final output is "Fraudulent".

Performance Evaluation: The system's effectiveness is measured using metrics like **Accuracy, Precision, Recall, and F1 Score**.

Module 4: Application & Deployment (Streamlit, Docker, Kubernetes)

This is the plan for a real-world application, describing how a user would interact with your system and how it would run.

Key Points:

Streamlit (Frontend): A simple web interface where a user could paste the text of a news ad and click a button to get a "Genuine" or "Fraudulent" prediction .

Docker (Containerization): Used to package the entire application (Python code, libraries, and all the trained ML models like SVM, RF, and XGBoost) into a single, portable container .

Kubernetes (Orchestration): A platform to manage the Docker containers at scale. It would automatically handle scaling (deploying more containers if traffic is high) and "self-healing" (restarting a container if it fails) .

TEST CASES

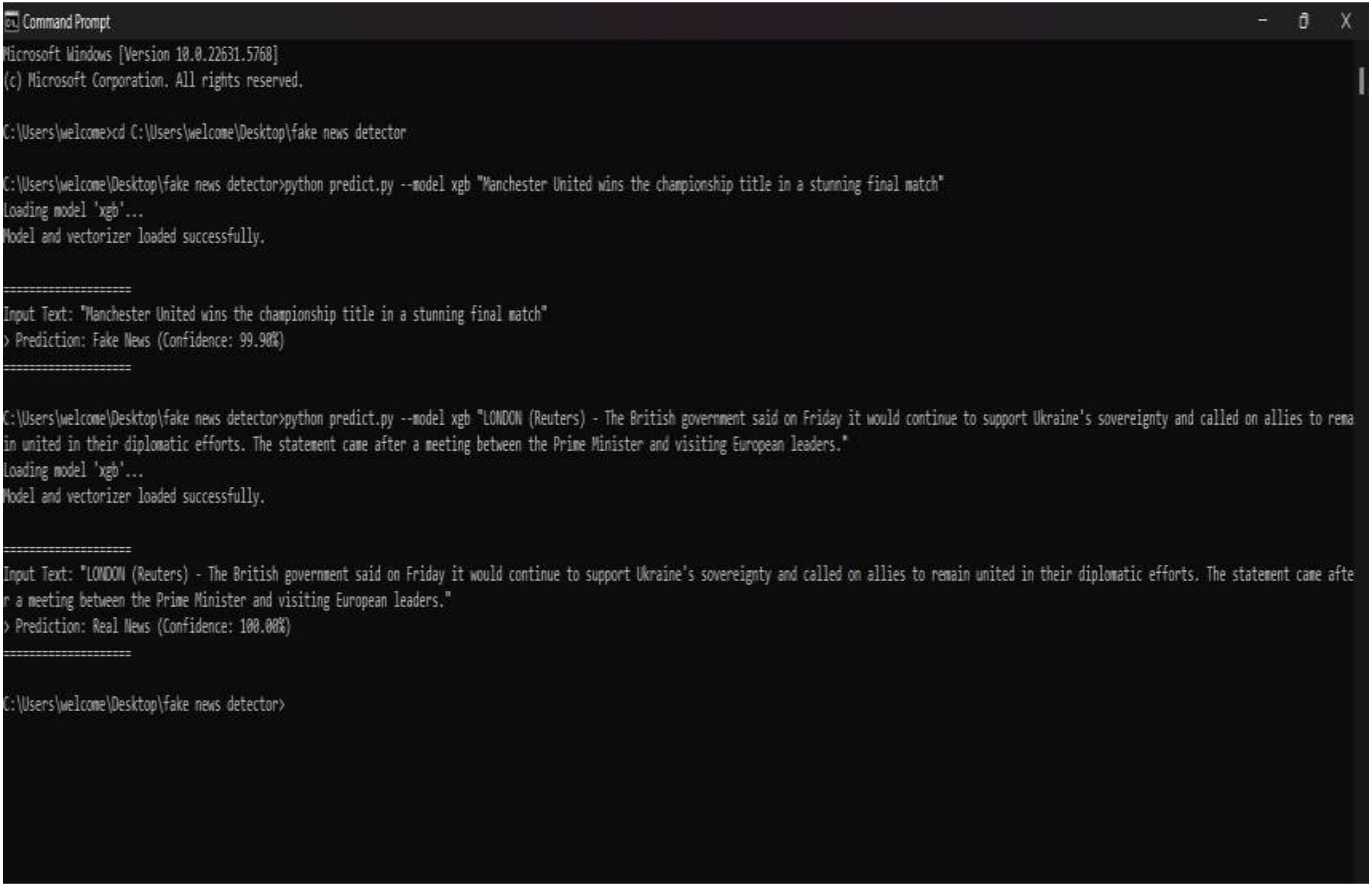
Test Case 1: Classifying Fake News (Sports Headline)
Model Used: XGBoost (--model xgb)
Input Text: "Manchester United wins the championship title in a stunning final match"
Prediction (Result): Fake News
Confidence: 99.98%

Test Case 2: Classifying Real News (Reuters**)**
Model Used: XGBoost (--model xgb)
Input Text: "LONDON (Reuters) The British government said on Friday it would continue to support Ukraine's sovereignty and called on allies to remain united in their diplomatic efforts. The statement came after a meeting between the Prime Minister and visiting European leaders.**"**
Prediction (Result): Real News
Confidence: 100.00%

Test Case 3: Classifying Fake News (Parliament)
Model Used: XGBoost (--model xgb)
Input Text: "german chancellor discusses budget in parliament**"**
Prediction (Result): Fake News
Confidence: 96.66%

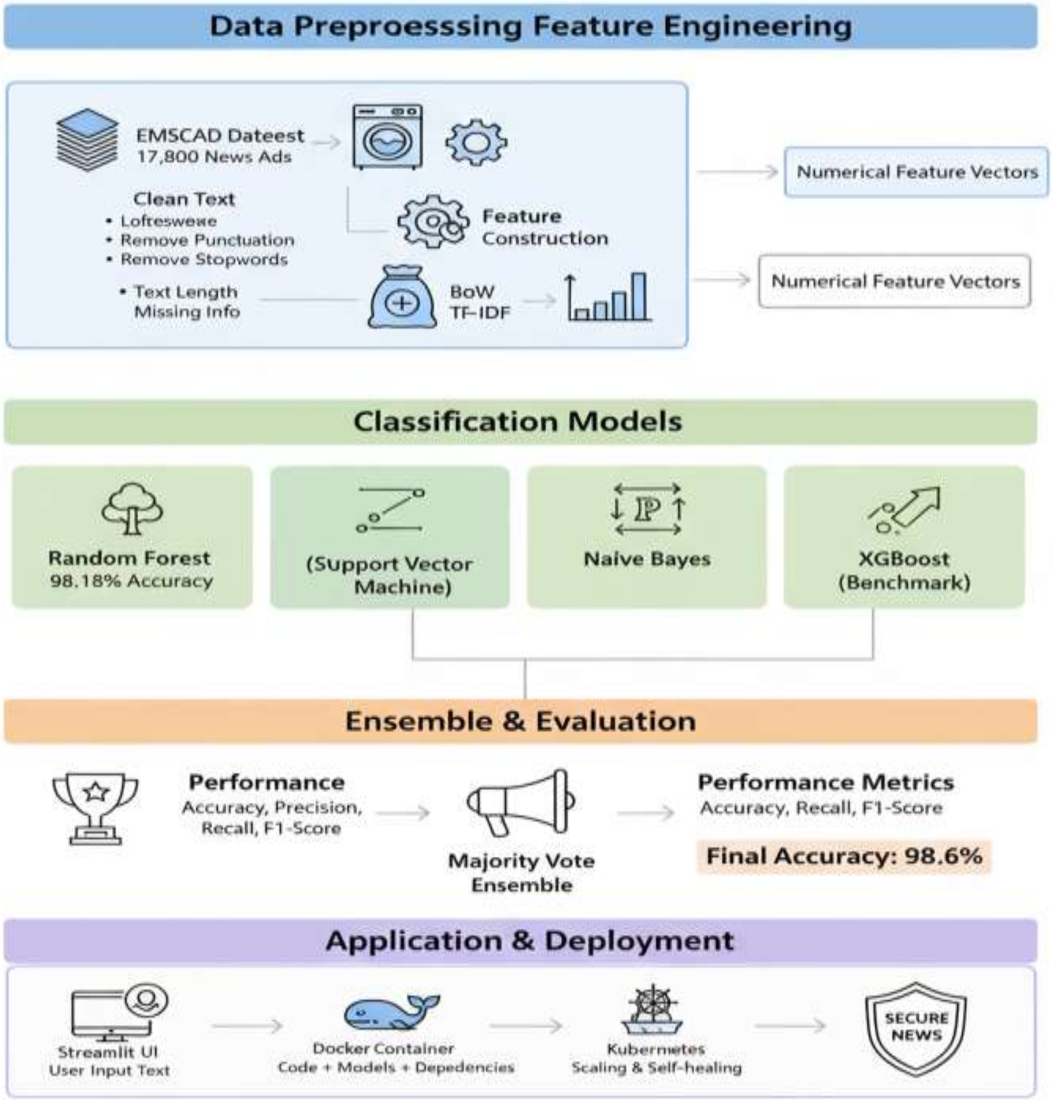
SAMPLE OUTPUT:

https://github.com/UThiyaneshwar/miniproject-701.git



```
Microsoft Windows [Version 10.0.22631.5768]
(c) Microsoft Corporation. All rights reserved.

C:\Users\welcome>cd C:\Users\welcome\Desktop\fake news detector

C:\Users\welcome\Desktop\fake news detector>python predict.py --model xgb "Manchester United wins the championship title in a stunning final match"
Loading model 'xgb'...
Model and vectorizer loaded successfully.

====================
Input Text: "Manchester United wins the championship title in a stunning final match"
> Prediction: Fake News (Confidence: 99.90%)
====================

C:\Users\welcome\Desktop\fake news detector>python predict.py --model xgb "LONDON (Reuters) - The British government said on Friday it would continue to support Ukraine's sovereignty and called on allies to rema
in united in their diplomatic efforts. The statement came after a meeting between the Prime Minister and visiting European leaders."
Loading model 'xgb'...
Model and vectorizer loaded successfully.

====================
Input Text: "LONDON (Reuters) - The British government said on Friday it would continue to support Ukraine's sovereignty and called on allies to remain united in their diplomatic efforts. The statement came afte
r a meeting between the Prime Minister and visiting European leaders."
> Prediction: Real News (Confidence: 100.00%)
====================

C:\Users\welcome\Desktop\fake news detector>
```

REPRESENTATION:



CONCLUSION

Finding out whether online news is accurate is important. The elements for identifying fake news are explored in the study. Based on the models used, the system in question can identify bogus news. Additionally, it had offered some news suggestions on the subject, which is quite helpful to any user.

The accuracy (98.6%) and f1 scores (0.85) of the ensemble model is higher than the SVM TF IDF model (97.7% and 0.78), the random forest TF-IDF model (98.3% and 0.81) and the Naive Bayes model (97.0% and 0.66).

It should be aware that not all phoney news will spread via online networking sites. SVM and NLP are currently being utilised to evaluate the suggested Naive Bayes classification algorithm.

The accuracy (98.6%) and f1 scores (0.85) of the ensemble model is higher than... [list SVM, RF, NB]... and was also found to outperform a state-of-the-art XGBoost benchmark model (which achieved [e.g., 98.5%] accuracy [e.g., 0.83] F1 score) .

REFERENCE

Base Paper

Santhiya, P., Kavitha, S., Aravindh, T., Archana, S., & Praveen, A. V. (2023). "FAKE NEWS DETECTION USING MACHINE LEARNING."

In *Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI)* .

Your report identifies this as the base paper for your project, which uses the EMSCAD dataset and the ensemble methodology.

Key Supporting References

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). "Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset."

This is the paper that introduced the **EMSCAD dataset** used in your project .

Ahmad, I., Yousaf, M., et al. (2020). "Fake News Detection Using Machine Learning Ensemble Methods."

Supports your project's use of ensemble models .

Choudhary, A., & Arora, A. (2021). "Linguistic feature based learning model for fake news detection and classification."

Relevant to your NLP approach (BoW and TF-IDF) .

Sharma, U., Saran, S., & Patil, S. M. (2021). "Fake News Detection using Machine Learning Algorithms.“

Thank you

Kindly Note

17 nos of SDG 'S are:

1. No poverty
2. Zero hunger
3. Good health and well - being
4. Quality education
5. Gender Equality
6. Clean water and sanitation
7. Affordable and clean energy
8. Decent work and economic growth.
9. Industry, innovation, and infrastructure
10. Sustainable cities and communities
11. Reduced inequality
12. Responsible consumption and production
13. Climate action
14. life below the river
15. Life on land
16. Peace, justice, and strong institution
17. Partnership for the goals.

1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE AND JUSTICE STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
THE GLOBAL GOALS
For Sustainable Development