SERVE-SMARTHACKATHON

INDIAN INSTITUTE OF TECHNOLOGY (BANARAS HINDU UNIVERSITY)

REPORT

TEAM NAME-RAJS161.ANSHRAJ

TEAM MEMBER-

1)ANSH SINGH

2)ADITYA KARN

3)ADARSH DUBEY

Methodology

1)Data Preprocessing:

- Tokenization and removal of stop words.
- Text vectorization using TF-IDF (Term Frequency-Inverse Document Frequency).
- Sentiment analysis to identify emotional biases.
- Feature extraction such as title length, text length, keyword density, and readability.

2)Feature Engineering:

- **Title Length**: Fake news often has exaggerated or sensational headlines.
- **Text Length**: Analyzing verbosity to detect content intended to obscure facts.
- **Keyword Density**: Examining frequent use of sensational keywords like "shocking" or "breaking."
- **Linguistic Features**: Grammar quality, punctuation usage, and readability.

3)Model Training:

- **Algorithm Used**: Random Forest Classifier.
- **Training Process**: The dataset was split into training and testing sets (80%-20% split).
- **Hyperparameters**:
    - Number of Trees: 100
    - Max Depth: Tuned based on cross-validation.

- For deep learning trials the following configuration was used:
    - **Model Architecture**: Sequential model with fully connected Dense layers.
    - **Optimizer**: Adam.
    - **Loss Function**: Binary Cross-Entropy.
    - **Number of Epochs**: 20.
    - **Batch Size**: 32.

4)Evaluation Metrics:

- **Accuracy**: Proportion of correctly classified instances.
- **Precision**: Ability to correctly identify positive instances among predicted positives.
- **Recall**: Ability to identify all actual positives.
- **F1 Score**: Harmonic mean of precision and recall.
- **AUC-ROC**: Measures the model's ability to distinguish between classes.

Conclusion

The project successfully built an efficient fake news detection model with an accuracy of 97.43% and an AUC-ROC score of 0.986, indicating excellent classification performance. Key takeaways include:

- **Features like title length, keyword density, and sentiment analysis significantly enhance model performance.**
- **The Random Forest Classifier proved effective for this classification task.**
- **Deep learning trials achieved promising results but required additional computational resources.**

Future Scope:

- **Incorporate more diverse datasets to improve generalizability.**
- **Explore advanced deep learning techniques like BERT for better contextual understanding.**
- **Integrate real-time prediction capabilities for broader applications.**

```
Accuracy: 0.9743
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.97      0.98      3137
           1       0.97      0.97      0.97      2863

    accuracy                           0.97      6000
   macro avg       0.97      0.97      0.97      6000
weighted avg       0.97      0.97      0.97      6000
```