**📘 Project Title:**

**A Comparative Analysis of Traditional and Deep Learning Approaches for Document Classification Using the 20 Newsgroups Dataset**

**🧠 Project Summary:**

In this research-driven project, I conducted a comprehensive comparative analysis between **traditional machine learning models** and **advanced deep learning architectures** for document classification using the **20 Newsgroups dataset**, a well-known benchmark in Natural Language Processing (NLP).

The project aimed to evaluate and improve the performance of multiple models in classifying unstructured text data into 20 predefined categories, addressing challenges such as class imbalance, noisy data, and high dimensionality. I developed and optimized models ranging from **Naive Bayes** and **Support Vector Machines (SVM)** to **Long Short-Term Memory (LSTM)**, **Convolutional Neural Networks (CNN)**, and **hybrid models** combining BiLSTM, CNN, GRU, and **attention mechanisms**.

**🏆 Key Achievements:**

* ⚙️ **Developed and evaluated 6+ classification models**, from baseline ML (Naive Bayes, SVM) to deep learning (LSTM, CNN) and hybrid attention-based models.
* 📈 **Achieved highest classification accuracy of 85.02%** and a macro-averaged **ROC-AUC of 0.9881** using a **BiLSTM with Attention mechanism**, outperforming all other models.
* 📊 Applied robust evaluation metrics (Accuracy, Precision, Recall, F1-Score, PR-AUC, ROC-AUC) to analyze model effectiveness and generalization capability.
* 🧩 **Implemented advanced optimization techniques** such as early stopping, learning rate schedulers, gradient clipping, and regularization to prevent overfitting.
* 📚 Pre-processed over 18,000 documents with tokenization, lemmatization, stop word removal, and used both **TF-IDF and GloVe embeddings** for feature representation.
* 🛠️ Built a **CNN-GRU-Attention hybrid model**, integrating sequential and local text features to boost classification in challenging classes.
* 📉 Identified limitations in traditional models’ ability to capture contextual relationships and validated the superior performance of deep learning on large-scale, complex text data.

**💡 Skills & Competencies Demonstrated:**

**🔬 Data Science & NLP**

* Text preprocessing, tokenization, lemmatization
* Feature engineering with TF-IDF, GloVe embeddings
* Handling imbalanced datasets and noisy text

**🤖 Machine Learning & Deep Learning**

* Traditional models: Naive Bayes, SVM
* Deep models: LSTM, CNN, GRU
* Advanced architectures: BiLSTM with Attention, CNN-GRU with Gradient Clipping
* Hyperparameter tuning, model fine-tuning, and architecture design

**📐 Evaluation & Optimization**

* Model evaluation with precision, recall, F1, PR-AUC, ROC-AUC
* Training with early stopping, learning rate scheduling, and dropout regularization

**🛠️ Technical Tools**

* **Python, TensorFlow/Keras, Scikit-learn, Numpy, Pandas, Matplotlib**
* Experience with Jupyter Notebooks, visualizations, and model performance analysis

**🌐 Outcome & Impact:**

This project not only deepened my understanding of text classification pipelines but also highlighted the scalability and robustness of hybrid deep learning models for real-world NLP tasks. It demonstrates my ability to bridge academic research with practical machine learning implementation — skills critical for roles in **data science**, **machine learning engineering**, or **AI research**.