Project Status Report – Group 21

1. **Project Title**: Detection of Fraudulent Job Postings With Machine Learning

2. **Problem Statement:**
   - Fraudulent Job listings have alarmingly increased in parallel with the exponential growth of online job platforms. These postings lead to financial and emotional damage and are often difficult to identify manually due to subtle language cues.
   - Our Project solves this problem by building an automated machine learning-based classifier to distinguish between real and fake job listings based on the text content provided in job descriptions and related fields. Our hypothesis is that certain linguistic patterns found in fraudulent job listings can be identified by machine learning models and captured by textual feature extraction.

3. **Modifications Based on Proposal Feedback:**
   - Added Logistic Regression as a traditional baseline model to compare against deep learning models.
   - Removed CNN, which is less suited for text classification and not central to our problem.
   - As BERT, BiLSTM require high computational power and GPU resources we considered using DistilBERT.
   - We currently use LSTM with GloVe for the deep learning baseline and will explore DistilBERT in next phase.

4. **Methodology and Experimental Plan:**
   - False job advertisements frequently have identifiable language indicators, such ambiguous job descriptions, exaggerated benefits, or lack of accurate corporate information. One of the most obvious signs of fraud are probably features taken from the specifications like title, and description areas.

   - As part of the preprocessing, the text was cleaned up by deleting special characters and HTML tags, converting it to lowercase, and tokenizing it. In order to maintain meaningful word forms, stopwords were eliminated and lemmatization was carried out using POS tagging. The main input for our analysis was the combination of the requirements and the job description. Additionally, missing values and class imbalance were addressed during the exploratory data analysis phase.

   - Following the preprocessing phase, three models were implemented for classification: Logistic Regression and XGBoost Classifier using TF-IDF features, and an LSTM model leveraging pre-trained GloVe embeddings for capturing contextual word representations. Additionally, for more sophisticated context representations, we intend to evaluate transformer-based models (DistilBERT).

5. **Tools and Frameworks:**
    - **Dataset**: Fake Job Postings Dataset from Kaggle (17,880 samples-800 are labled as fake)
    - **Languages and Platforms**: Python, Google Colab
    - **Libraries/Frameworks**: Scikit-learn, TensorFlow/Keras, NLTK, XGBoost, Matplotlib
    - **Embeddings**: Pre-trained GloVe vectors (100-dimensional)

6. **Current Status and Results:**

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	96.89%	1.00	35.84%	52.77%
XGBoost	98.01%	97.22%	60.69%	74.73%
LSTM + GloVe	95.16%	0.00	0.00	0.00

Issues identified:

- Despite obtaining an overall accuracy of 95.16%, the LSTM model was unable to identify any fraudulent job ads, giving a precision and recall of 0.00 for the fraud class.

- The model always predicts the majority class to minimize loss, mainly because of the dataset's extreme class imbalance, where real job ads dominate (~95%). It had no incentive to learn patterns linked to fraudulent instances because neither class weighting nor oversampling was used.

- This insight highlights the need for class balancing techniques, such as weighted loss functions or data resampling, which we plan to incorporate in the final phase.

7. **Brief Plan for the Remaining Month**

   To address limitations and finalize the project, we will:

- Implement class weighting and oversampling to handle imbalance.
- Integrate focal loss in deep learning models to penalize overconfident predictions.
- Extend experimentation with DistilBERT for advanced contextual understanding of text.
- Add auxiliary features such as company profile and benefits to enrich inputs.
- Perform extensive hyperparameter tuning, validation, and visualizations (e.g., learning curves, confusion matrices).

____________________________THANK YOU______________________________

Group Number 21

Members :- Sai Pavan P, Madhu Sree K, Venkata Sai Chandra V