Functional Document

Fake Job Post Detection using Machine Learning

1. Introduction

Employment scams are a significant issue in online recruitment, where fraudulent job postings aim to exploit job seekers, often by demanding money or sensitive personal information. The aim of the proposed system is to detect such fraudulent postings using machine learning (ML) classification techniques, ensuring legitimate opportunities are highlighted while scams are flagged.

2. Key Components

2.1. Dataset

- **Source**: Kaggle's "Real or Fake Job Posting Prediction" dataset.
- **Description**:
  - Comprises 17,880 job postings.
  - Attributes include job title, description, location, salary range, company profile, and fraudulent labels (target variable).
- **Preprocessing Steps:**
  - Missing Value Removal: Deletes records with incomplete or irrelevant data.
  - Stop-Word Elimination: Removes filler words like "the," "and," "or."
  - Irrelevant Attribute Removal: Filters out non-contributory columns.
  - Categorical Encoding: Converts categorical data into numerical format for ML compatibility.
  - Balanced Dataset Preparation: Ensures equal representation of fraudulent and legitimate job posts.

2.2. Machine Learning Classifiers

- Single Classifiers
  - Naive Bayes:
    - A probabilistic classifier using Bayes' Theorem.
    - Assumes feature independence, making it suitable for high-dimensional data.
    - Performs well under simple, structured data scenarios.
  - Multi-Layer Perceptron
    - A type of neural network with multiple hidden layers.
    - Uses backpropagation for optimization.
    - In the framework, configured with 5 hidden layers of sizes: 128, 64, 32, 16, and 8.
  - K-Nearest Neighbor
    - Lazy learning algorithm that classifies based on the proximity of data points in feature space.
    - Optimal performance observed for k=5k = 5k=5.
  - Decision Tree:
    - A tree-structured algorithm where nodes represent features, branches denote conditions, and leaves represent class labels.
    - Highly interpretable and competitive for spam or scam detection tasks.
- Ensemble Classifiers
  - Ensemble methods combine predictions from multiple models to enhance accuracy.
    - **Random Forest**:
      - Consists of multiple decision trees, each trained on random subsets of data.
      - Predicts the final class by majority voting from individual trees.
    - **AdaBoost**:
      - Combines weak learners into a strong learner by reweighting data points based on misclassifications.
      - Effective in improving classifier accuracy for unbalanced datasets.

- **Gradient Boosting**
  - Optimizes predictive performance by building trees sequentially.
  - Focuses on minimizing prediction loss at each step.

2.3. . Performance Metrics

- Accuracy
  - Proportion of correctly classified instances out of all instances.
- F1-Score
  - Harmonic mean of precision and recall, balancing false positives and false negatives.
- Cohen's Kappa Score:
  - Measures inter-rater agreement, adjusted for chance agreement.
  - High values indicate robust classifier agreement.
- Mean Squared Error:
  - Measures the average squared differences between predicted and true values.
  - Lower MSE indicates better model performance.

3. Proposed Methodology

3.1. Data Preparation

- Dataset is preprocessed to remove inconsistencies, followed by feature extraction to represent data numerically.

3.2. Model Training

- Models are trained using 80% of the dataset, and tested on the remaining 20%.
- Each classifier is configured with specific hyperparameters to enhance performance:
  - Random Forest: 500 estimators.
  - AdaBoost and Gradient Boost: 500 iterations.
  - KNN: k=5k = 5k=5.

3.3. Model Evaluation

- Single classifiers like Decision Tree and MLP are tested independently.

- Ensemble classifiers (Random Forest, AdaBoost, Gradient Boost) are evaluated and compared.

3.4. Final Selection

- Random Forest emerged as the best-performing model
  - **Accuracy**: 98.27%
  - **F1-Score**: 0.97
  - **Cohen's Kappa Score**: 0.74
  - **MSE**: 0.02

4. Comparison of Classifiers

4.1. Single Classifier Performance:

- Decision Tree and MLP show high accuracy (~97%) and F1-scores (~0.97).
- Naive Bayes has the lowest performance due to the dataset's complexity and feature dependencies.

4.2. Ensemble Classifier Performance:

- Random Forest outperforms all models with the highest accuracy and Cohen's Kappa score.
- AdaBoost and Gradient Boost also achieve competitive results.

5. Advantages of Random Forest

5.1. High accuracy and robustness due to multiple decision trees.

5.2. Effective handling of large, unbalanced datasets.

5.3. Low error rates (MSE of 0.02).

6. Applications

6.1. Job Portal Integration:

- Real-time classification of job postings.
- Automatic flagging of suspicious posts for review.

6.2. Employer Validation:

- Helps job seekers evaluate the legitimacy of companies.

6.3. Data Security:

- Reduces exposure to fraudulent practices, enhancing user trust.

7. Conclusion

The study demonstrates that machine learning, particularly ensemble methods like Random Forest, is effective in detecting fake job postings. With 98.27% accuracy, the proposed framework offers a reliable solution to tackle online recruitment fraud.