

Fake News Detection Using Machine Learning and NLP

1. Introduction and Aim

The rapid spread of misinformation on social media and digital platforms has created a global challenge in distinguishing real news from fake content. Traditional fact-checking methods are slow and labor-intensive. Therefore, this project aims to develop an AI-powered system that automatically detects fake news based on text analysis using machine learning (ML) and natural language processing (NLP).

Objectives:

- To design and train a classification model that distinguishes fake news from real news using linguistic and semantic features.
- To compare traditional ML models (e.g., SVM, Logistic Regression) with advanced deep learning models (e.g., LSTM, BERT).
- To build an interpretable system capable of explaining why a news article is classified as fake.

2. Related Work

Early works mainly relied on text preprocessing (tokenization, stopword removal, stemming) and feature extraction techniques such as TF-IDF or n-grams. Models like SVM, Logistic Regression, and Naïve Bayes achieved accuracies between 85–95% (Ahmed et al., 2017; Shu et al., 2020). Later, deep learning methods such as CNN and LSTM improved contextual understanding, while transformer-based models like BERT and FakeBERT achieved state-of-the-art performance (Kaliyar et al., 2021).

3. Techniques and Methods

1. Data Preprocessing: Clean and normalize text using tokenization, lemmatization, and stopword removal.
2. Feature Extraction: Represent text numerically using TF-IDF, Word2Vec, or BERT embeddings.
3. Model Training: Apply classifiers such as SVM, Random Forest, LSTM, and BERT.
4. Evaluation Metrics: Accuracy, Precision, Recall, and F1-score.

4. Data Sources

The following datasets will be used for model training and testing:

- Fake News Detection Dataset (Kaggle): 20,000+ English news articles labeled as real or fake. <https://www.kaggle.com/c/fake-news/data>
- LIAR Dataset: 12,000 short political statements with truth labels. <https://www.cs.ucsb.edu/~william/lie-dataset.html>
- FakeNewsNet: Repository combining news content, social context, and user data. <https://github.com/KaiDMML/FakeNewsNet>

5. Proposed Contribution

The proposed approach integrates traditional ML and deep learning:

1. Lightweight model (TF-IDF + SVM) for fast classification.
2. Deep model (BERT) for complex or ambiguous cases.
3. Explainability layer (e.g., LIME) to show which words influenced the decision.
4. Inclusion of Arabic or bilingual data to address multilingual limitations.

6. Conclusion

Fake news detection remains an evolving challenge in AI and NLP. By combining ML, transformers, and explainable AI, this project aims to deliver a practical and interpretable system that efficiently identifies misinformation and supports digital trust.

7. References

- Ahmed, H., Traore, I., & Saad, S. (2017). Detecting opinion spams and fake news using text classification. *Expert Systems with Applications*, 80, 67–79.
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Dynamic Information. *AAAI*.
- Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80(8), 11765–11780.
- Zhou, X., & Zafarani, R. (2020). A Survey of Fake News Detection: Data and Methods. *ACM Computing Surveys*, 53(5), 1–40.

8. Reference Summary Table

The following table summarizes the research institutions, techniques used, and publication sources related to the project.

Research Institution / Authors	Techniques Used	Data Source / Reference Link	Year
Ahmed, H., Traore, I., & Saad, S. (University of Victoria, Canada)	TF-IDF, Naïve Bayes, SVM for text classification	Expert Systems with Applications, 80, 67–79	2017
Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (Arizona State University)	FakeNewsNet dataset, social context + content analysis using ML	AAAI Conference Paper – FakeNewsNet Dataset Repository	2020
Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (Indian Institute of Technology)	FakeBERT model – BERT embeddings + CNN layers for semantic detection	Multimedia Tools and Applications, 80(8), 11765–11780	2021
Zhou, X., & Zafarani, R. (Syracuse University)	Survey on fake news datasets and detection methods (ML, DL, hybrid)	ACM Computing Surveys, 53(5), 1–40	2020

Employee Performance Evaluation Using Machine Learning

1. Introduction and Aim

Employee performance evaluation is a crucial process in organizations to assess workforce productivity, identify training needs, and support data-driven human resource decisions. Traditional methods are often subjective, time-consuming, and prone to bias. This project aims to develop an automated system that evaluates employee performance based on measurable factors using machine learning (ML) and data analytics techniques.

Objectives:

- To build a predictive model that accurately assesses employee performance.
- To identify the most influential factors affecting employee productivity.
- To assist management in making objective, data-driven decisions regarding promotions, rewards, and training.

2. Related Work

Several studies have explored the use of machine learning in performance evaluation. Early models used regression techniques to predict employee ratings based on KPI metrics such as working hours, experience, and project outcomes. Later works integrated classification algorithms (e.g., Random Forest, Decision Trees, and SVM) to categorize employees into performance levels such as 'Excellent', 'Average', or 'Needs Improvement' (Parmar et al., 2019; Mehta et al., 2021). Recently, deep learning models and hybrid techniques combining HR analytics and sentiment analysis of employee feedback have shown promising results in predicting engagement and performance trends.

3. Techniques and Methods

1. Data Preprocessing: Handle missing data, normalize continuous variables, and encode categorical values.
2. Feature Engineering: Extract features such as attendance rate, task completion time, peer feedback, and training hours.
3. Model Training: Implement ML models such as Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks.
4. Evaluation Metrics: Accuracy, F1-score, Mean Absolute Error (MAE), and R^2 Score.

4. Data Sources

Possible datasets for experimentation include:

- Kaggle: HR Analytics - Employee Performance Dataset
(<https://www.kaggle.com/datasets/rhuebner/human-resources-data>)
- IBM HR Analytics Employee Attrition & Performance Dataset
(<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>)
- Company internal HR records (if available) for real-world validation.

5. Proposed Contribution

This project proposes a hybrid system that combines performance metrics, attendance data, and employee feedback using ML algorithms. A multi-level model will be developed to

predict performance categories and suggest improvement areas. Additionally, an explainability layer (e.g., SHAP values) will be integrated to visualize how each factor influences the final score, promoting fairness and transparency.

6. Conclusion

By leveraging ML and AI-driven analytics, employee performance evaluation becomes more objective, efficient, and actionable. The proposed system provides HR departments with a reliable decision-support tool to identify high performers, reduce bias, and improve overall organizational productivity.

7. References

Parmar, D., Kataria, R., & Patel, V. (2019). Predicting employee performance using machine learning techniques. International Journal of Computer Applications, 975(8887), 1–5.

Mehta, A., & Rana, M. (2021). HR analytics for employee performance evaluation using Random Forest. International Journal of Advanced Research in Computer Science, 12(3), 22–28.

Mishra, R., & Kumar, S. (2020). Predictive analysis of employee performance using deep learning. Journal of Data Science and Analytics, 8(2), 45–54.

Kaggle HR Datasets: <https://www.kaggle.com/>

8. Reference Summary Table

The following table summarizes the key research sources, techniques, and publication details related to this project.

Research Institution / Authors	Techniques Used	Data Source / Reference Link	Year
Parmar, D., Kataria, R., & Patel, V. (India)	Regression and Decision Trees for predicting employee performance	International Journal of Computer Applications, 975(8887)	2019
Mehta, A., & Rana, M. (India)	Random Forest, HR Analytics Framework for classification	International Journal of Advanced Research in Computer Science, 12(3)	2021
Mishra, R., & Kumar, S. (India)	Deep Learning and ANN for employee performance prediction	Journal of Data Science and Analytics, 8(2)	2020
Kaggle HR Analytics Dataset (Public Repository)	Real-world HR performance data used for ML experiments	https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset	2022