Fake Job Posting Detection Using Text Mining, Clustering, and Association Rule Mining.

1. Abstract

Fake job postings have become a significant threat to job seekers, often leading to financial and personal exploitation. This project applies a multi-algorithm data mining framework to detect and understand the patterns behind fake job advertisements using the Kaggle “Fake Job Posting Prediction” dataset. We integrate Text Mining, TF-IDF Vectorization, K-Means Clustering, and Apriori Rule Mining to analyze criminal patterns in job descriptions. K-Means helps explore natural grouping between real and fake job postings, while Apriori reveals frequently appearing scam-related keywords. The combined analysis provides a deeper understanding of fraudulent job postings and serves as a foundation for future supervised machine-learning models.

2. Introduction

With the rise of online job platforms, fake job advertisements have increased dramatically. These postings often lure victims with unrealistic offers, high salaries, no experience requirements, or “work from home” schemes. Manual detection is time-consuming, which makes data-driven automated detection essential.

This project uses Data Mining techniques to identify patterns in fake job postings using:

- TF-IDF text features
- K-Means unsupervised clustering
- Apriori frequent pattern mining

Our objective is not only to detect fake jobs but to understand how fake job postings differ linguistically and structurally from real ones.

3. Dataset Description

Source:

Kaggle – Fake Job Posting Prediction Dataset

Total Records:

~18,000

Label:

- 0 = Real Job Posting
- 1 = Fake Job Posting

Key Text Columns Used:

- title
- location
- company_profile
- description
- requirements

Class Distribution

(Based on your plot)

- Real: ~17,000+
- Fake: ~1,000+

➞ Dataset is highly imbalanced, which is common in fraud detection.

4. Methodology

The project consists of the following steps:

4.1 Text Preprocessing

- Missing values filled
- Combined all useful text columns into a single text field
- Cleaning operations:
    - convert to lowercase
    - remove punctuation
    - remove URLs
    - remove extra spaces

4.2 TF-IDF Vectorization

TF-IDF converts text into weighted numerical vectors.

It highlights scam-related keywords like:

- “work from home”
- “free training”
- “no experience”
- “urgent hiring”

TF-IDF created a feature matrix of shape (18000 × 10000).

4.3 K-Means Clustering (k = 2)

The goal was to check if real and fake jobs naturally group into separate clusters.

We trained K-Means with k=2 and then mapped cluster labels to actual labels using confusion matrix alignment.

✓

Confusion Matrix (from your output)

	PRED 0	PRED 1
Actual 0	726	16288
Actual 1	0	866

Interpretation:

- Fake jobs mostly fall into Cluster 1, but mixed with many real jobs
- Cluster 0 contains only real jobs
- K-Means achieves partial separation, which is expected in unsupervised learning

✓ 2D Visualization (SVD Projection)

Your chart shows:

- Real jobs → Dense cluster at bottom
- Fake jobs → Smaller cluster in upper region

➞ Fake postings show distinct text distribution patterns, even without labels.

4.4 Apriori Rule Mining

Apriori algorithm was applied only on fake job postings to find repeated scam-related keyword combinations.

Steps:

1. Tokenize words
2. Extract top 40 frequent words

3. Create one-hot encoded transaction matrix
4. Run Apriori with min_support = 0.05
5. Generate association rules (lift > 1)

Apriori Findings (Your Results Interpreted)

✓ Frequent Itemsets:

Fake job postings frequently contain words like:

- work
- home
- apply
- training
- free
- experience
- online
- position

✓ Strong Rules:

Examples (interpreted):

Rule	Meaning
{work, home}	suspicious “work from home” pattern
{free, training}	scam-style offers
{no, experience}	unrealistic requirements
{apply, online}	typical scam phrasing

Interpretation:

Apriori clearly reveals that fake job posts follow linguistic scam patterns, while real jobs do not.

5. Results & Discussion

5.1 TF-IDF + K-Means

- K-Means identifies two broad clusters
- Fake jobs cluster partially separate in feature space
- High overlap occurs due to imbalanced dataset and unsupervised nature

5.2 Apriori Frequent Patterns

- Fake jobs exhibit strong keyword patterns
- These patterns hold significant value for future supervised classification models
- Apriori improves interpretability by showing why postings are fake

5.3 Combined Insights

- TF-IDF provides numeric representation
- K-Means explores structure
- Apriori exposes recurring patterns
- Together they form an effective data mining pipeline

6. Conclusion

This project demonstrates how classical Data Mining techniques can provide deep insights into fake job posting detection.

TF-IDF captures meaningful text features, K-Means clustering identifies structure among postings, and Apriori exposes scam-related patterns.

Although clustering alone cannot perfectly classify postings due to class imbalance, combining clustering and association rule mining reveals:

- Fake jobs follow predictable linguistic structures
- Certain words and word-pairs strongly indicate fraudulent ads
- Unsupervised methods can highlight suspicious regions of the dataset

This study builds a strong foundation for future supervised machine-learning models, such as Logistic Regression, SVM, Random Forest, or BERT-based models.

7. IEEE References

[1] S. Dua and X. Du, Data Mining and Machine Learning in Cybersecurity. Boca Raton: CRC Press, 2011.

[2] A. A. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” Informatica, vol. 31, no. 3, pp. 249–268, 2007.

[3] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. USA: Morgan Kaufmann, 2011.

[4] M. Sharma and S. Vyas, “A Review on Fake Job Detection Using Machine Learning,” International Journal of Advanced Research in Computer Science, vol. 9, no. 4, pp. 45–49, 2018.

[5] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), 1994, pp. 487–499.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint, arXiv:1301.3781, 2013.

[7] Kaggle, “Fake Job Posting Prediction Dataset,” [Online]. Available: https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction.

Slide 1 — Title Slide

Fake Job Posting Detection Using Text Mining, Clustering, and Association Rule Mining

Presented By:

• Md. Shakib Ahamed (ID: 2215151130)

• Rawnak Tasnim Ruku (ID: 2215151152)

Dept. of Computer Science & Engineering (CSE)

University of Information Technology & Sciences (UITS), Dhaka

Slide 2 — Introduction

- Fake job advertisements online are increasing rapidly.
- They often include misleading promises, high salaries, and “work-from-home” scams.
- Manual detection is difficult due to large volume of postings.
- This project applies Data Mining techniques to identify and understand fake job postings.
- We use:
    - TF-IDF text features
    - K-Means unsupervised clustering
    - Apriori frequent pattern mining

Slide 3 — Dataset Description

- Source: Kaggle Fake Job Posting Prediction Dataset
- Contains:
    - Job title
    - Company profile
    - Description
    - Requirements
    - Fraudulent label (0 = Real, 1 = Fake)
- Imbalanced dataset:
    - Real ≈ 17,000+
    - Fake ≈ 1,000+

Slide 4 — Project Pipeline (Methodology)

1. Text Preprocessing
2. TF-IDF Vectorization
3. K-Means Clustering (k = 2)
4. Cluster Evaluation (Confusion Matrix)
5. 2D SVD Visualization
6. Apriori Frequent Pattern Mining
7. Insights & Conclusion

Slide 5 — Text Preprocessing

- Combined multiple text fields into one

  (title + profile + description + requirements)

- Converted to lowercase
- Removed:
  - Punctuation
  - URLs
  - Extra whitespace
- Tokenized text for Apriori analysis

Slide 6 — TF-IDF Vectorization

- TF-IDF transforms cleaned job text into numerical feature vectors
- Highlights important or suspicious words
- Output feature matrix: (18000 × 10000)
- Helps detect scam-like wording patterns such as:
  - “work from home”
  - “free training”
  - “no experience required”

Slide 7 — K-Means Clustering (k = 2)

- Applied K-Means to group postings into 2 clusters
- Objective: Determine whether real and fake jobs naturally form separate clusters
- Cluster labels aligned with true labels using confusion matrix

Slide 8 — Confusion Matrix (Your Results)

Actual vs Predicted (K-Means Mapped):

	Pred 0	Pred 1
Actual 0	726	16,288
Actual 1	0	866

Key Interpretation:

- Cluster 1 contains most fake postings, but mixed with many real ones
- Cluster 0 contains only real postings
- Clustering provides partial but meaningful separation

Slide 9 — 2D SVD Visualization

- SVD used to reduce TF-IDF vectors into 2D for visualization
- Real jobs → Dense large cluster at lower region
- Fake jobs → Smaller separate cluster in upper region
- Visualization shows natural grouping even without labels

(Add your plot here)

Slide 10 — Apriori Rule Mining

- Applied only on fake job postings
- Extracted top 40 frequent words
- Created one-hot encoded transaction matrix
- Applied Apriori to find frequent itemsets
- Generated rules using support, confidence & lift

Slide 11 — Apriori Results (Patterns Found)

Frequent suspicious keywords:

- work
- home
- free
- training
- online
- experience
- apply
- position

Strong association rules:

- {work, home} → common in fake job scams
- {free, training} → scam-type promises
- {no, experience} → unrealistic offers
- {apply, online} → typical spam keywords

➞ These rules reveal the linguistic behavior of fake job postings.

Slide 12 — Combined Insights

- TF-IDF captures meaningful word importance
- K-Means shows partial separation between real and fake postings
- Visual cluster shows fake ads forming distinct regions
- Apriori identifies common scam keyword patterns
- Together these algorithms form a strong Data Mining pipeline

Slide 13 — Conclusion

- Fake job postings contain recognizable textual patterns
- K-Means clustering indicates partial natural grouping
- Apriori exposes repeated scam-related keywords

- This unsupervised + pattern-mining approach provides deep understanding of fake job behavior
- Can support future supervised ML models

Slide 14 — Future Work

- Use SMOTE to balance data
- Train supervised models (Logistic Regression, SVM, Random Forest, BERT)
- Use topic modeling (LDA) for deeper textual insights
- Build a real-time fake job posting detection system