Task 1

An analysis of approximately 18,000 job postings—highlighted significant linguistic and structural differences between real and fake job advertisements. This study employed three primary methods: n-gram frequency analysis, missing data comparison, and Part-of-Speech tagging to discern patterns indicative of fraudulent postings.

N-gram

Real Job Postings (Charts 3 & 5): Common bigrams such as "customer service" and "business development," along with trigrams like "monthly cost living," suggest that these postings represent well-structured and professional roles. These phrases indicate clear job functions, responsibilities, and expectations typically associated with legitimate employment opportunities. By providing specific information regarding job requirements, compensation, and corporate duties, these postings exhibit transparency and attention to detail, which helps candidates assess the role and its expectations comprehensively.

Fake Job Postings (Charts 4 & 6): Phrases such as "oil gas" and "work home" in bigrams, alongside trigrams like "solutions global provider," point to vague and potentially deceptive job descriptions. Although industries such as oil and gas are legitimate, they might be targeted in fraudulent postings because of the perceived allure of high-paying, remote roles. Additionally, the use of generic or ambiguous terms, coupled with poorly formatted text (e.g., "nbsp nbsp"), signals a lack of precision and care in these postings.

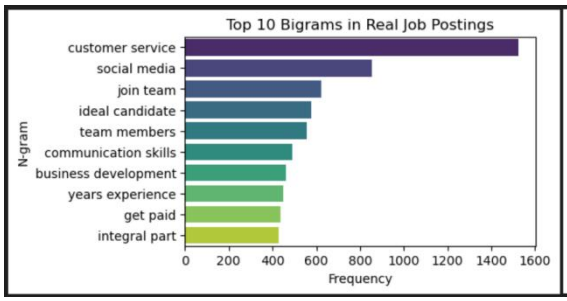


Chart 3

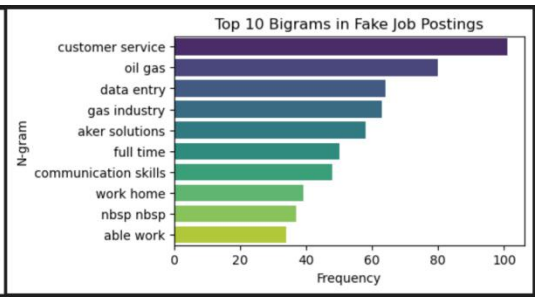


Chart 4

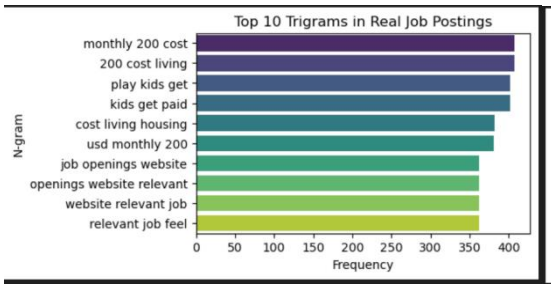


Chart 5

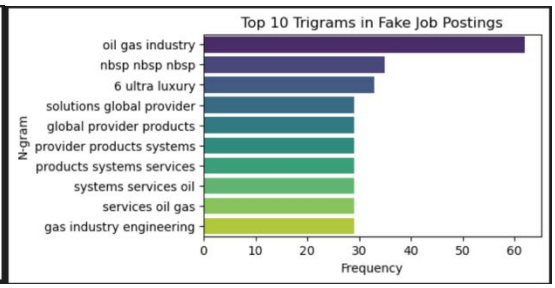


Chart 6

Missing Data Comparison

The missing data comparison further distinguished real from fake postings.

Real job postings (chart 7) had higher instances of omitted information in fields like salary

range, department, and required education—possibly because established companies assume certain details are understood or deem them non-essential at the initial application stage. Fake postings (chart 8), on the other hand, showed significant missing data in critical fields such as company profile, requirements, and location. This absence suggests that fake job postings are more likely to lack critical information, either because the details cannot be provided or because the ambiguity is intended to prevent candidates from recognizing the posting as fraudulent.

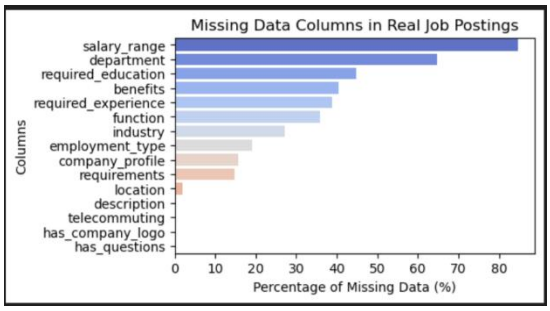


Chart 7

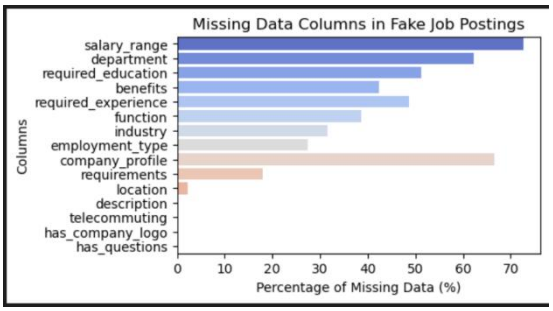


Chart 8

POS Tagging

The POS tagging analysis reveals that both real and fake job postings follow similar patterns in the use of nouns, adjectives, and verbs. However, real job postings consistently have a higher word count, especially in fields like company profile and requirements. This suggests that real postings provide more detailed and specific information about the job and company. In contrast, fake job postings tend to be more concise and ambiguous, offering fewer concrete details, which may indicate fraudulent intent.

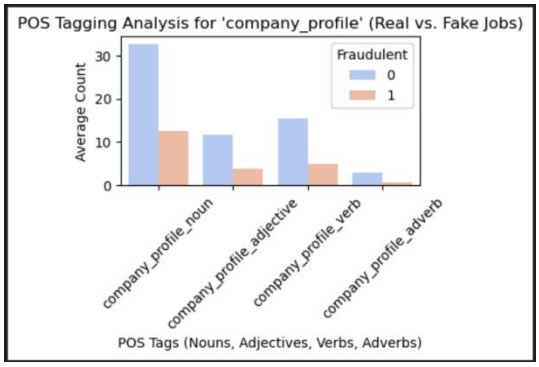


Chart 9

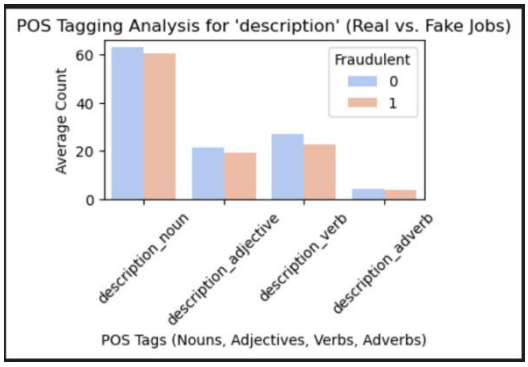
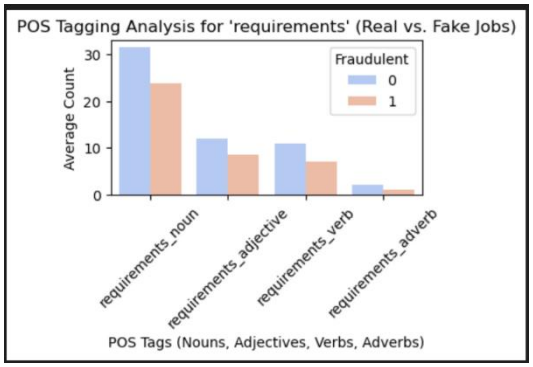


Chart 10



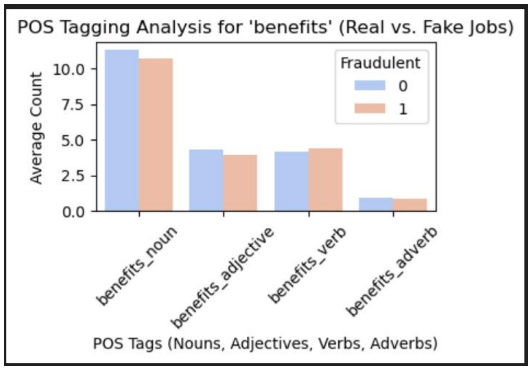


Chart 11 Chart 12

Conclusion

In conclusion, real job postings are characterized by specificity, detailed language, and transparency, focusing on concrete job aspects even when some non-critical fields are incomplete. They provide essential information that enables applicants to make informed decisions. Conversely, fake job postings tend to use vague, promotional language, omit crucial details, and exhibit structural deficiencies like missing key information and poor formatting.

Task 2a: Classification (TF-IDF)

The classification process utilized a logistic regression model with the regularization parameter C, which was optimized through a grid search combined with cross-validation. Grid search systematically explores all possible hyperparameter combinations to identify the optimal model configuration. To ensure robust results and reduce the risk of overfitting, 5-fold cross-validation was applied, averaging performance across multiple validation splits.

A wide range of C values, from 0.001 to 5000, was tested to examine various levels of regularization. Smaller C values correspond to stronger regularization, helping to prevent overfitting, while larger C values reduce regularization, allowing the model greater flexibility in fitting the data. This broad spectrum ensured that different regularization strengths were thoroughly explored. The complete set of tested C values included:

[0.001, 0.01, 0.1, 1, 5, 10, 50, 100, 200, 500, 750, 1000, 2000, 5000].

After initial testing, the values C=30 and C=200 was identified as optimal for the educational classification and fraud detection tasks, respectively. Further fine-tuning was then conducted around these values to ensure the stability and reliability of the results.

education level classification		Fraudulent job prediction	
C	**mean accuracy**	**C**	**mean accuracy**
20	0.9066 (+/-0.0053)	150	0.9775 (+/-0.0017)
22	0.9068 (+/-0.0041)	160	0.9781 (+/-0.0013)
24	0.9068 (+/-0.0046)	170	0.9779 (+/-0.0019)
26	0.9069 (+/-0.0045)	180	0.9777 (+/-0.0017)
28	0.9071 (+/-0.0056)	190	0.9778 (+/-0.0019)
30	0.9073 (+/-0.0050)	200	0.9783 (+/-0.0014)
32	0.9073 (+/-0.0055)	210	0.9776 (+/-0.0021)
34	0.9077 (+/-0.0060)	220	0.9779 (+/-0.0013)
36	0.9069 (+/-0.0048)	230	0.9775 (+/-0.0017)
38	0.9075 (+/-0.0059)	240	0.9775 (+/-0.0014)
40	0.9069 (+/-0.0064)	250	0.9770 (+/-0.0023)

Table 1

Where C = 34 gives the highest mean accuracy for the education level classification and C =

200 gives the highest mean accuracy for the fraudulent job prediction (table 1).

Results

Metrics	Education level classification	Fraudulent job classification (fraudulent=1)
Best accuracy on validation set	0.9073	0.9783
Best C value	34	200
Precision	0.8672	0.9043
Recall	0.7236	0.6012
F1-score	0.7595	0.7222
Accuracy on test set	0.9183	0.9776

Table 2

According to result (table 2), the best-performing C value for the educational classification task was 34, resulting in a validation accuracy of 0.9073. The model achieved an accuracy of 0.9183 on the test set, demonstrating solid generalization performance. While the accuracy was high, I decided to further evaluate precision, recall, and F1 score to ensure that the model's performance was balanced across all classes, and that it didn’t suffer from issues such as misclassification of minority classes. High precision and recall values for the three educational classes underscore the model's effectiveness at predicting educational levels based on job descriptions. The F1 score, which balances precision and recall, was 0.7595, indicating that the model performed well in maintaining a balance between false positives and false negatives.

For the fraudulent job detection task, the best C value was 200, with a validation accuracy of 0.9783. The model achieved a high-test accuracy of 0.9776. The precision of 0.9043 highlights the model's ability to minimize false positives (correctly identifying real job postings). However, the recall of 0.6012 suggests that some fraudulent job postings were missed. The F1 score of 0.7222 reflects a reasonable trade-off between precision and recall, though improvements in recall could enhance the model's ability to detect more fraudulent postings.

Conclusion:

Grid search, combined with 5-fold cross-validation, was instrumental in determining optimal C values for both tasks. A wide range of C values was explored to balance overfitting and underfitting, with the best-performing C values of 34 for educational level classification and 200 for fraudulent job detection. The test set accuracies were 0.9183 and 0.9776, respectively.

Task 2b: Word2Vec

In this task, three key hyper-parameters were selected and tuned: vector size, window size, and epochs. The ranges for these parameters were selected based on a balance between computational efficiency and the need to capture sufficient linguistic context (table 3).

Range of hyper-parameters			
Vector size	100	200	300
windows	2	5	10
epochs	5	10	/

Table 3

Vector size: 100, 200, and 300 dimensions were chosen for the embeddings. These sizes are commonly used in NLP tasks, with 100 being lightweight for fast computation and 300 often providing deeper semantic information. The 200 dimension was included as a midpoint to test if it could strike an optimal balance between performance and complexity.

Window size: A range of 2, 5, and 10 was chosen to explore different context levels. Smaller window sizes (e.g., 2) focus on very local context, which can be useful for capturing specific relationships, while larger windows (e.g., 10) provide broader contextual understanding, which may be essential for job descriptions (i.e., large surrounding text).

Epochs: The number of epochs was set to 5 and 10. These values were selected to provide sufficient training iterations without excessive overfitting, balancing computational cost and performance.

C-value: the range of C remains the same as previous task, to ensure flexibility and prevent overfitting.

[0.001, 0.01, 0.1, 1, 5, 10, 50, 100, 200, 500, 750, 1000, 2000, 5000].

Results

Metrics	Education level classification	Fraudulent job classification (fraudulent=1)
Best accuracy on validation set	0.9562	0.9591
Best C value	1	10
Best vector size:	200	200
Best window size:	10	10
Best epochs:	10	10
Precision	0.7709	0.5976
Recall	0.6272	0.2832
F1-score	0.6400	0.3843
Accuracy on test set	0.8816	0.9561

Table 4

After grid search and cross-validation, the best-performing settings for the education level classification task were listed in Table 4. The consistency of vector size and window size

across tasks highlights that these settings effectively capture the linguistic patterns in job descriptions, while the difference in C values indicates varying regularization needs.

Test Set Accuracy Comparison

The education level classification task achieved a validation accuracy of 0.9562 and a test set accuracy of 0.8816. While this performance seems relatively high, the F1-score of 0.6400 indicates that there is significant room for improvement. The F1-score suggests that the model is struggling to balance the precision and recall, meaning that the Word2Vec-based model is not robust enough in predicting educational levels accurately across all test instances.

For the fraudulent job classification, despite the validation accuracy of 0.9591 and test accuracy of 0.9561, the model performed poorly in terms of F1-score (0.3843). This low F1-score indicates a severe challenge in detecting fraudulent job postings due to the class imbalance, where fraudulent jobs are significantly outnumbered by real ones. However, even after testing a wider range of the hyperparameters and set the balance dataset, the improvements in F1-score remained marginal, which means Word2vec is not appropriate for this task.

When compared to the performance of TF-IDF, Word2Vec shows clear limitations on both tasks. This is likely due to TF-IDF's focus on the statistical relevance of specific words tied closely to the classification labels. By giving more weight to distinctive terms, such as 'Master's Degree' or 'fraud,' TF-IDF enhances the model's ability to identify critical features that directly influence classification outcomes. In contrast, Word2Vec captures broader semantic relationships, which may dilute its effectiveness in tasks that rely on precise keyword distinctions for accuracy.

Task 3: Transformer

Architecture of the Classifier

The transformer-based model used for job description classification tasks is built using PyTorch. The architecture begins with an embedding layer that maps the input tokens into a high-dimensional space.

```
JobClassifier(
  (token_embeddings): Embedding(9227, 300, padding_idx=0)
  (dropout): Dropout(p=0.5, inplace=False)
  (transformer_encoder): TransformerEncoder(
    (layers): ModuleList(
      (0-1): 2 x TransformerEncoderLayer(
        (self_attn): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=300, out_features=300, bias=True)
        )
        (linear1): Linear(in_features=300, out_features=512, bias=True)
        (dropout): Dropout(p=0.5, inplace=False)
        (linear2): Linear(in_features=512, out_features=300, bias=True)
        (norm1): LayerNorm((300,), eps=1e-05, elementwise_affine=True)
        (norm2): LayerNorm((300,), eps=1e-05, elementwise_affine=True)
        (dropout1): Dropout(p=0.5, inplace=False)
        (dropout2): Dropout(p=0.5, inplace=False)
      )
    )
  )
  (linear): Linear(in_features=300, out_features=1, bias=True)
)
```

The model uses two transformer encoder layers, each equipped with multi-head self-attention

mechanisms and feed-forward networks. The hidden dimensionality of both the attention mechanism and feed-forward network is set to 300. Dropout layers with a rate of 0.5 are applied after the embedding and attention layers to prevent overfitting. A positional encoding layer ensures that the model captures the sequential and contextual information in the input sequence. Finally, a linear layer with a single output unit is used to compute the logits for binary classification.

Accuracy on the Test Set of Transformer

For the two classification tasks, the model achieved the following results (Table 5):

Metrics	**Education level classification**	**Fraudulent job classification (fraudulent=1)**
Validation accuracy	0.8011	0.9810
Precision	0.8304	0.8835
Recall	0.8768	0.5260
F1-score	0.8522	0.6594
Accuracy on test set	0.8768	0.9737

Table 5

These results reflect the model's ability to balance precision and recall effectively, as seen from the relatively high F1 scores for both tasks.

Comparison of Classification Performance

For both tasks, the transformer-based model was compared against TF-IDF and Word2Vec approaches in terms of accuracy and F1 score (Table 6):

	Education Level Classification		**Fraudulent Job Detection**	
Methods	Accuracy	F1 score	Accuracy	F1 score
Transformer	0.8768	0.8522	0.9737	0.6594
TF-IDF	0.9183	0.7595	0.9776	0.7222
Word2Vec	0.8816	0.6400	0.9561	0.3843

Table 6

When comparing the transformer model to TF-IDF and Word2Vec, we observe differences in both accuracy and F1 score. While TF-IDF achieved the highest accuracy in both tasks, the transformer model outperformed it in terms of F1 score for the education level classification. This indicates that while TF-IDF can efficiently detect relevant keywords, transformers excel at maintaining a balance between precision and recall. The contextual awareness in the transformer model leads to better handling of complex language patterns, which can explain its higher F1 score despite slightly lower accuracy.

On the other hand, Word2Vec, though competitive in accuracy, performed poorly in F1 score, particularly for fraudulent job detection. The lower F1 score suggests that Word2Vec struggles to balance precision and recall in imbalanced datasets, where fraudulent postings are significantly outnumbered by legitimate ones. This indicates a limitation of Word2Vec in capturing fine-grained relationships between tokens in job descriptions.

Advantages and Limitations of the Three Approaches

Transformer: Captures complex relationships and context between words effectively, leading to a balanced classification performance. However, it is computationally expensive and requires significant time and resources for training.
TF-IDF: Excels in keyword-based classification by leveraging statistical word frequency but lacks the capability to capture contextual relationships between words, limiting its effectiveness in more nuanced tasks.
Word2Vec: Efficient at capturing semantic relationships and is computationally lightweight. However, it struggles with class imbalance and achieving a good balance between precision and recall, making it less suitable for tasks like fraudulent job detection.

Task 4a: DistilBERT

	Fraudulent job detection		Education level classification	
Type	Origin	Fine Tune	Origin	Fine Tune
Precision	0.8349	0.9491	0.5212	0.8749
Recall	0.8909	0.9522	0.7105	0.8876
F1-Score	0.8573	0.9491	0.5959	0.8764

Table 7

The fine-tuned DistilBERT outperformed the original model in both tasks (Table 7). For fraudulent job detection, the fine-tuned model achieved significantly higher precision (0.9491 vs. 0.8349) and recall (0.9522 vs. 0.8909), indicating better accuracy and fewer false positives. In the education level classification, the fine-tuned model also demonstrated substantial improvements in precision (0.8749 vs. 0.5212) and F1-score (0.8764 vs. 0.5959). Fine-tuning allows the model to adapt to specific task data, resulting in higher performance, especially for multi-class classification. In contrast, the original model struggled due to its more general training.

Task 4b:

In Task 4b, zero-shot classification was employed using GPT-o1 and Gemini-1.5 models to predict educational levels in job postings. A subset of 10 job postings was selected from rows 36-47, excluding rows 39 and 46, as they lacked educational information. GPT-o1 achieved 100% accuracy across all 10 job postings, while Gemini-1.5 correctly predicted 9 out of 10 cases.
The key advantage of zero-shot learning lies in its flexibility. Both models were able to classify text without requiring additional fine-tuning, making them suitable for unstructured tasks like job posting classification. GPT-o1's superior accuracy can be attributed to its broader training data and better ability to capture context, while Gemini-1.5, though effective, may have had limitations due to differences in model architecture or the specific training data it was exposed to.
While zero-shot learning is highly effective in handling diverse tasks, it may face limitations when applied to domain-specific or more nuanced job postings where detailed context is

required. Despite this, the overall performance of both models demonstrates the power of zero-shot learning in predicting educational levels in job descriptions with minimal task-specific adaptation.

Job ID	Predicted Educational Level	Actual Required Educational Level	Correctness of Prediction
1	Bachelor's Degree	Bachelor's Degree	Correct
2	Bachelor's Degree	Bachelor's Degree	Correct
3	Bachelor's Degree	Bachelor's Degree	Correct
4	High School or equivalent	High School or equivalent	Correct
5	Bachelor's Degree	Bachelor's Degree	Correct
6	Bachelor's Degree	Bachelor's Degree	Correct
7	Bachelor's Degree	Bachelor's Degree	Correct
8	High School or equivalent	High School or equivalent	Correct
9	High School or equivalent	High School or equivalent	Correct
10	Bachelor's Degree	Bachelor's Degree	Correct

Table 8: GPT o1

Job Posting	Gemini 1.5 Prediction	Actual Result	Correct Prediction
1	Bachelor's Degree	Bachelor's Degree	Yes
2	Master's Degree	Bachelor's Degree	No
3	Bachelor's Degree	Bachelor's Degree	Yes
4	High School or equivalent	High School or equivalent	Yes
5	Bachelor's Degree	Bachelor's Degree	Yes
6	Bachelor's Degree	Bachelor's Degree	Yes
7	Bachelor's Degree	Bachelor's Degree	Yes
8	High School or equivalent	High School or equivalent	Yes
9	High School or equivalent	High School or equivalent	Yes
10	Bachelor's Degree	Bachelor's Degree	Yes

Table 9: Gemini 1.5

Task 5: Academic Integrity

Throughout the completion of Assignment 2, I focused on understanding the concepts behind data classification, applying techniques such as TF-IDF, Word2Vec, and models like Transformers and BERT. For coding, I consistently referred to lecture materials to grasp the theoretical concepts. However, when facing unfamiliar coding challenges—particularly when implementing a step-by-step Transformer model and distilBERT model—I used generative AI tools like GPT to assist in debugging and code generation. This approach streamlined my learning process, as it provided quicker solutions compared to traditional resources like StackOverflow. By analyzing the generated code step-by-step, I ensured that I fully understood its workings and annotated it with my own comments for self-learning.

While using GPT for coding assistance offered efficiency, I critically evaluated its use to ensure it complemented rather than replaced my learning. In contrast, for my academic writing and reports, I avoided using AI to generate content, opting instead to create my own analyses, charts, and explanations. I only utilized AI for grammar checking and language refinements to ensure clarity. This balance allowed me to uphold academic integrity while leveraging AI as a supportive learning tool without compromising my understanding or originality in the work.