Integrating Unsupervised Clustering with NLP for Online Job Fraud Detection

Inchul Yang
Vanderbilt University
inchul.yang@vanderbilt.edu

ABSTRACT

Fraudulent job postings pose significant risks to job seekers and undermine the trustworthiness of online job platforms. This study explores the use of natural language processing (NLP) and unsupervised learning techniques to detect and classify fraudulent job posts. The analysis applies K-means clustering to semantic embeddings derived from job posting texts to identify inherent patterns indicative of fraud. The findings show that one cluster heavily concentrates on fraudulent postings, while another is primarily associated with authentic listings. Logistic regression models further evaluate the effectiveness of cluster assignments as predictive features, both independently and when integrated with metadata and semantic embeddings. While cluster assignments improve models that combine metadata features, they do not enhance performance when combined with semantic embeddings. The embeddings intrinsically capture clustering information, suggesting their standalone use for fraud detection is preferred over adding clustering assignments. This study sheds light on the linguistic patterns of job scams and demonstrates how unsupervised techniques can expose semantic cues of fraud that enhance detection models, contributing to more secure online job markets.

Keywords

NLP, Semantic Embeddings, K-means clustering, Classification

1. INTRODUCTION

Fraudulent job postings not only waste the time and resources of job seekers but also jeopardize their personal information and online safety [1]. The impact of these deceptions reaches beyond the affected individuals, undermining the trustworthiness and functional integrity of job platforms. There is a clear need for an advanced method to systematically identify and classify online job scams. NLP presents a valuable tool to improve existing detection systems by delving into the semantic and linguistic nuances existing in job advertisements [2].

This research builds on prior studies that have employed machine-learning techniques to analyze and identify fraudulent job postings [3, 4, 5]. By examining semantic cues within the text of job posts, this study strives to distinguish authentic job posts from deceptive ones, thereby bolstering the security of job seekers and the dependability of online job markets. Moreover, this study aims to shed light on the linguistic tactics used by scammers and create an NLP-driven model that accurately detects fraudulent postings, thus providing both preventive and analytical benefits.

2. PURPOSE STATEMENT AND RESEARCH QUESTION

This study evaluates the effectiveness of unsupervised learning techniques for detecting fraudulent job postings through NLP analysis. Using spaCy embeddings for feature extraction and features derived from clustering analysis, the research explores the potential to distinguish authentic from fraudulent postings by enhancing the feature space. Additionally, it investigates how these unsupervised learning techniques, particularly clustering algorithms, can be integrated as features in classification models to improve performance [6, 7]. The anticipated outcomes aim to strengthen security protocols across job platforms and offer protection for both employers and job seekers against potential fraud.

This study is based on the following research questions:

1. How does clustering semantic embeddings from job postings help to identify distinctive patterns? What are the frequent words for each cluster, and how do they relate to authentic or fraudulent job postings?
2. How does the clustering assignment resulting from the analysis effectively classify job postings as authentic or fraudulent, both independently and in conjunction with other features?

3. LITERATURE REVIEW

The detection of fraudulent job postings has been an emerging research area in recent years. One of the pioneering works in this domain was by Vidros, Kolias, and Kambourakis (2016), where they introduced the concept of Online Recruitment Fraud (ORF) and highlighted the growing threat it poses to job seekers [8]. They pointed out that fraudsters were increasingly exploiting online job platforms to gather personal information from unsuspecting applicants.

In a follow-up study, Vidros, Kolias, Kambourakis, and Akoglu (2017) further explored the efficacy of machine learning classifiers in detecting fraudulent job advertisements. Their research applied a bag-of-words model and a binary rule-based feature set to a dataset of fraudulent job posts, yielding a 91% accuracy using a Random Forest classifier [9]. The significance of their findings was underscored by their decision to release the Employment Scam Aegean Dataset to the public, catalyzing subsequent investigations in the field.

Building on this foundation, later studies have significantly extended the scope of this research. Mahbub and Pardede (2018) proposed a new feature space that includes contextual features not previously addressed [3]. This innovation has notably improved the detection performance of fraudulent job advertisements across multiple classifiers. Similarly, Alghamdi and Alharby (2019) combined rule-based features with advanced machine-learning techniques to create a more sophisticated model [10]. This research contributes a new detection model that utilizes two data mining algorithms: Support Vector Machine (SVM) for feature selection and Random Forest for classification.

Lal, Jaiswal, Sardana, Verma, Kaur, and Mourya (2019), Dutta and Bandyopadhyay (2020), Mehboob and Malik (2021), and Naudé, Adebayo, and Nanda (2023) experimented with various

classification algorithms, all striving to identify the most effective feature sets and refine the accuracy of fraudulent job post detection [5, 11, 12, 13]. This collective research effort has consistently utilized the EMSCAD dataset, reaffirming its value as a resource for ongoing academic inquiry into employment scams.

In summary, the literature demonstrates significant research efforts toward detecting fraudulent job postings, marked by the development of various machine learning models and the utilization of rule-based features, contextual information, and classification algorithms. However, despite these advances, there is a relative scarcity of research exploring the potential of unsupervised learning methods, such as clustering, to uncover inherent patterns indicative of fraud. The present study aims to address these gaps by leveraging NLP and unsupervised clustering techniques to enhance the feature space for more effective classification of fraudulent posts.

4. METHODS

4.1. Data

The dataset employed in this study is the Employment Scam Aegean Dataset (EMSCAD), which was released to the public by Vidros et al. (2017) [9]. It comprises 17,880 job descriptions collected between 2012 and 2014 in collaboration with the job advertisement platform Workable. The job postings within the dataset were annotated by specialized employees from Workable, adhering to company policies and detailed analytical standards. All personal information in the job advertisements was anonymized to comply with GDPR regulations, ensuring the protection of personal data. The full EMSCAD dataset is accessible at: https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction.

This analysis leverages comprehensive job descriptions and the associated metadata to explore patterns indicative of fraudulent activities. Additional details about the data used in this analysis, including associated code and methodologies, can be found at: https://github.com/icever/NLP_project_paper.

4.1.1. Labels

In the EMSCAD dataset, a subset of job descriptions is explicitly labeled as fraudulent, representing about 5% of the total, with 866 entries, while the remaining 17,014 are marked as non-fraudulent. This labeling provides a foundation for differentiating between authentic and fraudulent job postings.

4.1.2. Textual Features

The dataset is rich in textual data, encompassing 13 separate textual features as listed in Table 1. These features are crucial for NLP applications in this analysis.

4.1.3. Metadata Features

The dataset also includes three specific metadata features that provide additional contextual information about each job posting, as outlined in Table 2.

4.1.4. Semantic Document Embeddings

Document embeddings were generated using a pre-trained spaCy model to encapsulate the semantic meanings of the textual content listed in Table 1. Data preprocessing was performed to facilitate the generation of high-quality 300-dimensional semantic document embeddings, which captured the essential semantic features of each job posting in a dense vector representation. The comprehensive data wrangling process laid the foundation for the effective application of unsupervised and supervised techniques in the study.

Table 1. Textual features

Name	Description
title	title of the job posting
location	geographic location of the job posting
department	corporate department
salary_range	approximate salary range
company_profile	brief description of the company
description	detailed description of the job posting
requirements	requirements listed for the vacancy
benefits	benefits offered
employment_type	type of employment
required_experience	required experience
required_education	required education
industry	industry
function	function performed

Table 2. Metadata features

Name	Description
telecommuting	1 for remote, 0 otherwise
has_no_company_logo	1 if no company logo is present, 0 otherwise
has_no_questions	1 if no screening questions are present, 0 otherwise

4.1.5. Data Preprocessing

Data preprocessing was undertaken to prepare the dataset for analysis. Given the significant class imbalance, with only 5% of the data labeled as fraudulent, an equal number of non-fraudulent job posts were randomly sampled to achieve a balanced representation. This step ensured that the subsequent analysis and modeling would not be biased by the skewed distribution.

Additionally, the textual data underwent thorough cleaning. The process began with data cleaning, where NaN values were replaced with blank spaces. Subsequently, all textual fields were merged into a single text column. Non-alphabetical characters and URLs were removed to refine the data, although punctuation was retained until after tokenization to allow the tokenizer to utilize these marks for accurate segmentation of the text.

4.2. Clustering

K-means clustering was applied to group similar job postings into the same clusters. This method ensures that data points within a cluster exhibit higher similarity to each other compared to those in different clusters. It is particularly useful for identifying inherent patterns in job postings, which may reveal common traits of fraudulent postings when contrasted with authentic ones.

For clustering, 300-dimensional spaCy embeddings were utilized as the features for each job posting. These embeddings effectively capture the semantic properties of the text and provide a foundation for clustering. The dataset was partitioned into K distinct, non-overlapping clusters, with each data point assigned to a single group. The optimal number of clusters, K, was determined through Silhouette Analysis, which helped confirm the coherence and separation of the clusters formed.

4.2.1. Silhouette Analysis

Silhouette Analysis was employed to determine the optimal number of clusters necessary for the cluster analysis. This technique is critical for evaluating the effectiveness of the clustering configuration by measuring how well each object lies within its cluster, a key to determining the right number of clusters.

$$silhouette\ score = \frac{(b-a)}{max(a,b)}$$

This analysis quantifies the quality of clustering by assessing how similar each data point is to its cluster compared to others. This similarity is calculated using the silhouette score, which reflects the degree of confidence in the clustering assignments. The score for each data point is computed based on two distances: the mean intra-cluster distance (a) and the mean nearest-cluster distance (b). The intra-cluster distance is the average distance of a data point from all other points in the same cluster, while the nearest-cluster distance is the average distance of the data point from the points in the nearest cluster.

4.2.2. Principal Component Analysis (PCA)

PCA was utilized in this study to reduce the embedding dimensionality from 300 to two. With the reduced dimensionality, it becomes feasible to visualize the data on a 2-dimensional plot, where each point represents a job posting and its placement is determined by the two principal components that explain the majority of the variance. This visualization is instrumental in identifying underlying patterns and clusters in the data, providing insights that are critical for subsequent clustering using K-means.

4.3. Classification

In this study, logistic regression models for classification were employed to evaluate the effectiveness of different feature combinations in predicting fraudulent job postings. These models were applied to understand how cluster assignments as features, alongside embeddings and metadata, influence the accuracy of fraud detection.

4.3.1. Model Construction

The classification models were developed to discern between authentic and fraudulent job postings, with the target variable defined as the job post's fraudulent status. The features employed in model construction, as listed in Table 3, were selected based on the premise that each provides distinct insights into the data. The integration of these distinct features is hypothesized to augment the model's capacity to accurately identify fraudulent job postings. To achieve this synergy, the features from each category will be concatenated, forming a comprehensive feature set that captures a wide spectrum of information.

In Table 3, Embeddings refer to the 300-dimensional vectors derived from a pre-trained spaCy model. Cluster Assignments are represented as one-hot encoded binary vectors, which result from the clustering analysis in this study. Metadata includes three features outlined in Table 2.

Table 3. Combinations of features

Model	Dim	Features
M1	4	Cluster assignments
M2	3	Metadata
M3	300	Embeddings
M4	7	Cluster assignments and Metadata
M5	303	Embeddings and Metadata
M6	304	Cluster assignments and Embeddings

Note: Dim is an abbreviation for Dimension.

4.3.2. One Hot Encoding

One hot encoding was employed to effectively use the categorical cluster assignments. It transforms each categorical feature into a series of binary variables, each representing a single cluster. For instance, if a data point is assigned to Cluster 0, then the variable for Cluster 0 is set to 1, while the variables for Clusters 1, 2, and 3 are set to 0. This method ensures that each cluster assignment is mutually exclusive and captures the unique groupings identified by the clustering algorithm.

4.3.3. Bootstrap Resampling

Bootstrap resampling was applied to estimate the distribution of accuracy statistics for each model presented in Table 3. By random sampling with replacement from the dataset, the samples were subsequently used to train the model and evaluate its performance. This process was repeated 1,000 times, resulting in a distribution of performance metrics from the bootstrap samples. This methodology allows for a more robust estimation of each model's accuracy by capturing the variability in performance across different samples.

4.3.4. Performance Evaluation Metrics

In this study, accuracy serves as the primary metric for evaluating the performance of the various classification models developed to detect fraudulent job postings. The decision to use accuracy is driven by its straightforward interpretability and effectiveness in providing a clear measure of the proportion of correct predictions made by a model. This metric is particularly useful when comparing the efficacy of different models, as it allows a direct comparison of how well each model performs across all predictions.

4.3.5. Statistical Analysis

Tukey's Honest Significant Difference (HSD) test was used to perform pairwise comparisons between the mean accuracies of different models obtained from bootstrap resampling. This statistical test identifies significant differences in group means. One research question examined whether combining cluster assignments with other features could additively enhance classification performance in detecting fraudulent job postings. By comparing models with different feature combinations, the analysis evaluated if integrating cluster assignments yielded superior accuracy over using features like metadata or embeddings alone. Tukey's HSD test enabled assessing the statistical significance of accuracy differences across models, shedding light on the additive predictive value of cluster assignment features.

5. RESULTS

5.1. Clustering

To determine the optimal number of clusters, factors such as the distribution and cohesion of the data points within them must be considered, in addition to the average silhouette score.

Figure 1 illustrates the silhouette scores for each data point across a range of cluster counts from 2 to 5. While clustering into two groups yields the highest average silhouette score, the distribution of data points across these clusters is not ideal, and there is a notable lack of cohesion within one of the clusters. Similar issues are seen in the three-cluster configuration. Furthermore, the five-cluster option results in negative silhouette scores for several clusters, indicating poor cluster formation. Therefore, clustering into four groups appears to be the most balanced solution among the configurations examined.

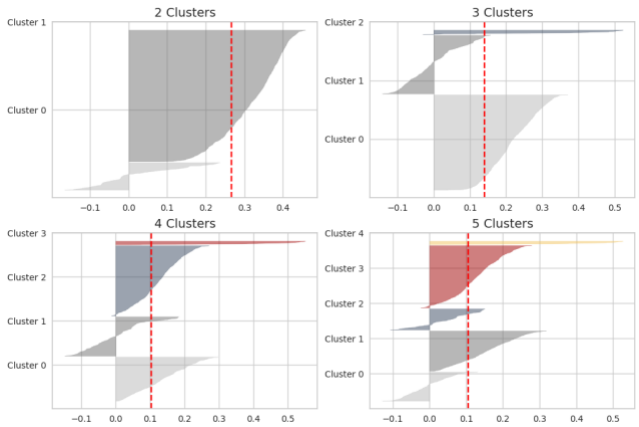

Figure 1. Silhouette plots for different numbers of clusters

Figure 2 demonstrates the clustering of job postings represented in a two-dimensional space derived from PCA. The scatter plot emphasizes a significant divide between clusters, with a prominent aggregation of fraudulent posts in the lower-right area, specifically within Cluster 3. In contrast, the upper-left area appears to be heavily composed of authentic job postings, predominantly associated with Cluster 0.

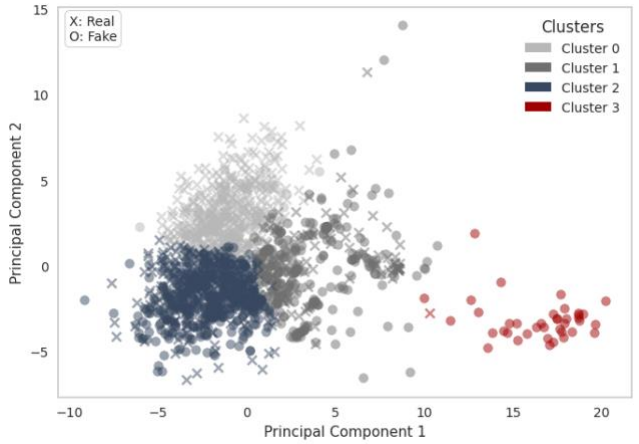

Figure 2. Cluster results visualized in two PCA dimensions

Complementing this, Figure 3 presents a bar chart that quantifies the distribution of authentic and fraudulent posts within the identified clusters. The contrast in Cluster 3 is evident, as it encompasses the largest proportion of fraudulent postings at 97.5%, whereas Cluster 0 exhibits a substantial majority of authentic postings, accounting for 85%. This marked presentation of data points suggests that clustering based on semantic embeddings can discern underlying patterns that could indicate features significant for classification purposes.

Table 4 details the top ten words associated with each cluster, as derived from the clustering analysis. It is particularly noteworthy that Cluster 3, which has a high portion of fraudulent job posts, is characterized by frequent terms such as 'time', '$', 'cash', and 'start'. This vocabulary suggests a pattern where deceptive postings often use language that promises quick financial rewards or emphasizes immediacy and flexibility.

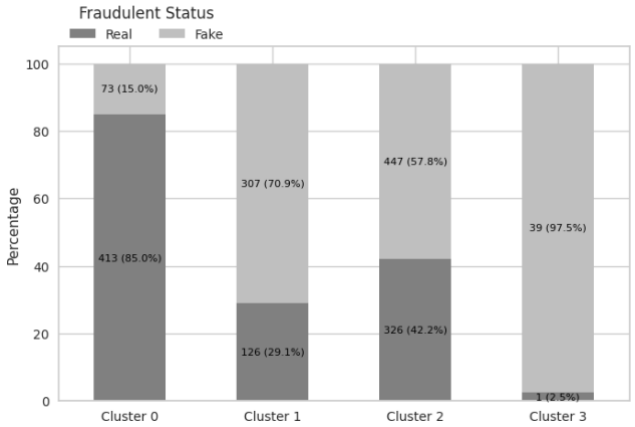

Figure 3. Distribution of real and fake posts across 4 clusters

In contrast, the words in Cluster 0, which include 'work', 'team', 'experience', and 'product', indicate authentic job postings. These terms are associated with professional roles and point to job descriptions that prioritize clear job responsibilities and qualifications. This linguistic distinction underscores that authentic job postings tend to communicate explicit expectations, particularly concerning business, or product management roles.

Table 4. Top 10 words for each cluster

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
1	work	work	work	time
2	team	time	experience	$
3	experience	experience	service	cash
4	product	amp	customer	start
5	company	service	business	require
6	time	job	team	day
7	design	position	time	experience
8	new	company	product	amp
9	technology	customer	project	free
10	business	entry	skill	work

5.2. Classification

The logistic regression analysis in Table 5 examines the impact of cluster assignments on the classification of job posts. The analysis is grounded in a dataset comprising 1,732 observations. The pseudo-R-squared value indicates that the model explains 16.9% of the variability in job post authenticity. This is not a particularly high value, but it does indicate that the cluster assignments capture some useful information for predicting fraud. Moreover, the very low Likelihood Ratio (LLR) p-value, at 1.31E-87, suggests that the model is highly significant.

In addition, each coefficient in the logistic regression output is statistically significant, with p-values falling below the 0.001 threshold. The magnitudes of the coefficients from Cluster 0 to Cluster 3 align with the pattern observed in Figures 2 and 3, where Cluster 3 is strongly associated with fraudulent posts, while Cluster 0 predominantly indicates authentic posts. These coefficients support the inference that certain cluster assignments are strongly associated with the probability of a job post being fraudulent. The results thereby validate the utility of cluster-based features in discerning the authenticity of job postings and provide evidence for their application in predictive models.

Table 5. Logistic regression results with cluster assignments on the prediction of fraudulent job posts

No. Observations				1732
Pseudo R-squared				0.1690
Log-Likelihood				-997.7
LLR p-value				1.31E-87
variable	coef.	stderr.	z	p
Cluster 0	-1.733	0.127	-13.649	0.000
Cluster 1	0.8906	0.106	8.417	0.000
Cluster 2	0.3157	0.073	4.334	0.000
Cluster 3	3.6636	1.013	3.617	0.000

Tables 6 and 7 present the accuracy distribution from bootstrap resampling and modeling outcomes, and pairwise Tukey's HSD test results, which demonstrate significant differences in accuracy across models. These findings underscore the substantial impact of feature selection on model performance.

The results in Table 6 reveal that Model M1, which solely relies on cluster assignments as features, achieves an accuracy of 69.6%. This finding, coupled with the significant coefficients observed in the logistic regression analysis in Table 5, provides strong evidence supporting the hypothesis that cluster assignments derived from semantic embeddings possess predictive value for identifying fraudulent job postings.

The research question further investigates how effectively clustering results can classify job postings when integrated with additional features. Models M4, which combines cluster assignments and metadata, and M6, which integrates cluster assignments with semantic embeddings, are pivotal for determining whether the synergy of combined features enhances classification accuracy.

Table 6. Accuracy from bootstrap resampling and modeling

Model	**Mean**	**Standard Deviation**
M1	69.6 %	2.5 %
M2	73.9 %	2.4 %
M3	90.6 %	1.6 %
M4	74.4 %	2.5 %
M5	91.5 %	1.5 %
M6	90.6 %	1.6 %

Table 7. Pairwise Tukey's HSD test results

Group1	**Group2**	**Mean Diff**	**P-adj**	**Reject**
M1	M2	0.042	0.00	TRUE
M1	M3	0.208	0.00	TRUE
M1	M4	0.048	0.00	TRUE
M1	M5	0.220	0.00	TRUE
M1	M6	0.209	0.00	TRUE
M2	M3	0.166	0.00	TRUE
M2	M4	0.006	0.00	TRUE
M2	M5	0.178	0.00	TRUE
M2	M6	0.167	0.00	TRUE
M3	M4	-0.160	0.00	TRUE
M3	M5	0.012	0.00	TRUE
M3	M6	0.001	0.99	FALSE
M4	M5	0.171	0.00	TRUE
M4	M6	0.161	0.00	TRUE
M5	M6	-0.011	0.00	TRUE

To evaluate the effectiveness of feature integration, Models M1, featuring cluster assignments, and M2, featuring metadata, were compared to M4, which uses a combination of the two features. As depicted in Figure 4, the violin plot shows that M4 exhibits higher mean accuracy compared to M1 and M2. This difference is statistically significant, as confirmed by the results in Table 7. These findings suggest that integrating cluster assignments with metadata can enhance model performance.

However, evaluating the effectiveness of feature integration for Model M6, which utilizes cluster assignments and semantic embeddings, yields differing results. This evaluation involves

comparisons with Model M1, which is based on cluster assignments, and Model M3, which utilizes semantic embeddings. As depicted in Figure 5, the violin plot indicates that M3 and M6 exhibit nearly identical mean accuracies. This difference is not statistically significant, as confirmed by the results in Table 7, suggesting that the integration of cluster assignments with semantic embeddings does not enhance model performance.

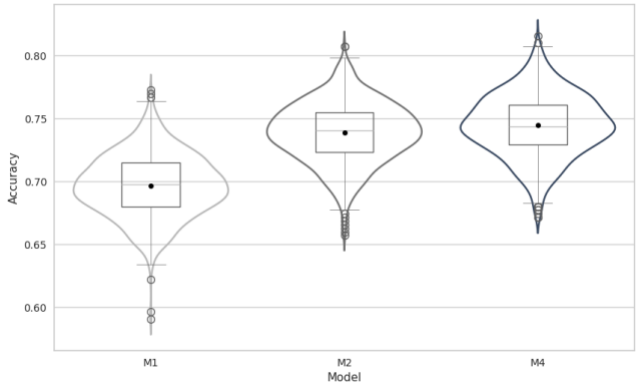

Figure 4. Comparison of accuracies: M1, M2, and M4

Note: M1 = Cluster assignments only, M2 = Metadata only, M4 = Combination of cluster assignments and metadata

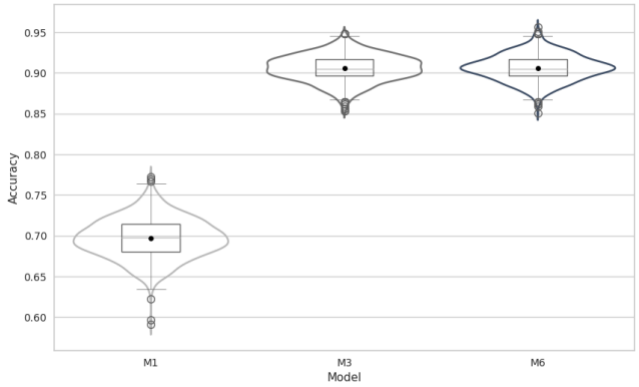

Figure 5. Comparison of accuracies: M1, M3, and M6

Note: M1 = Cluster assignments only, M3 = Embeddings only, M6 = Combination of cluster assignments and embeddings

The lack of performance improvement can be attributed to the fact that the clustering was derived from the embeddings and thus did not extract any new, distinguishable information beyond what was already captured within the embedding vectors. Furthermore, these findings suggest that the embedding vectors themselves already encapsulate the clustered features, making them robust for classification independently. Therefore, the added complexity of incorporating cluster information may not justify the minimal gains, making the use of embedding vectors alone a preferable option due to their simplicity and effectiveness.

6. CONCLUSION AND FUTURE WORK

This study has examined the effectiveness of clustering algorithms in analyzing job postings to uncover distinct patterns and investigate the relationship between cluster assignments and the likelihood of postings being authentic or fraudulent. The results indicate that a four-cluster configuration effectively groups similar embeddings from each job post. Notably, one specific cluster was strongly associated with fraudulent postings, while another was primarily associated with authentic listings. This clear distinction highlights the potential of using cluster assignments as features in a classification model. Moreover, the transformation of semantic embeddings of job postings into a two-dimensional space has been advantageous for visualizing and understanding the inherent patterns in the data.

To assess the predictive power of cluster assignments for classifying job posting authenticity, logistic regression analysis was performed. The results conclusively established a significant relationship between cluster assignments and the likelihood of a posting being fraudulent or authentic. The logistic regression model demonstrated a good fit to the data, explaining 16.9% of the variability in the data. Each cluster assignment coefficient was statistically significant, underscoring the model's robust predictive accuracy. Furthermore, the model achieved an accuracy of 69.6% when using only cluster assignments as features. These results provide compelling evidence that cluster assignments, derived from the semantic embeddings of job postings, inherently capture informative patterns that are valuable for detecting fraudulent listings. These findings strongly support the hypothesis that unsupervised clustering techniques can yield predictive features for enhancing fraud detection models in this domain.

The comparative analysis across models integrating different feature combinations underlines the potential for composite features to bolster classification performance. Specifically, the results indicate that incorporating clustering assignments can meaningfully augment existing feature sets, as evidenced by the improved accuracy observed when combined with metadata features. However, the integration of cluster assignments with semantic embeddings did not produce statistically significant improvements over using embeddings alone. This finding suggests that the high-dimensional embedding vectors intrinsically capture the salient information encapsulated by the clustering assignments, rendering the explicit inclusion of cluster assignments redundant. Consequently, the study's findings advocate for the use of semantic embeddings as a standalone feature, as they prove equally effective while circumventing the additional complexity introduced by integrating clustering assignments.

For future work, it is important to consider the potential imbalance in feature weighting that may have influenced the model construction and evaluation. Given that the embedding vectors consist of 300 dimensions while the cluster assignments are limited to only 4 dimensions, the embeddings likely exerted a disproportionate influence on the model's performance. To obtain a more accurate assessment of the impact of cluster assignments, it would be beneficial to balance the relative weighting of these features. This could involve either reducing the dimensionality of the embedding vectors or exploring techniques to increase the dimensional complexity of the cluster assignments. Additionally, rather than simply concatenating the cluster assignment features, alternative methods for combining these vectors, such as addition, averaging, or employing a gating mechanism, could be investigated to optimize their synergistic potential.

Furthermore, the current analysis solely relied on document embeddings derived from the textual content of job postings for clustering purposes. Future research should explore the incorporation of additional relevant features to generate more refined and informative clusters. By leveraging these enhanced cluster assignments in conjunction with the semantic embeddings, the classification models could potentially achieve improved performance in distinguishing fraudulent job postings from authentic ones.

7. REFERENCES

[1] Sharma, K., Parashar, D., Sangwan, A., Vengali, A., & Agrawal, G. (2023). Enhancing online job posting security: A big data approach to frau/d detection. *2023 IEEE International Carnahan Conference on Security Technology (ICCST)* (pp. 1-6). https://doi.org/10.1109/ICCST59048.2023.10474243

[2] Snidhuja, B., Anitha, B., Sowmya, A., & Srivalli, D. (2023). Prediction of fake job ad using NLP-based multilayer perceptron. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 14(1), 296-310. https://doi.org/10.17762/turcomat.v14i1.13533

[3] Mahbub, S., & Pardede, E. (2018). Using contextual features for online recruitment fraud detection. *Designing Digitalization (ISD2018 Proceedings)* (pp. 1-10). Lund, Sweden: Lund University. ISBN 978-91-7753-876-9.

[4] Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security*, 10(3), 155-176. https://doi.org/10.4236/jis.2019.103009.

[5] Naudé, M., Adebayo, K.J. & Nanda, R. A machine learning approach to detecting fraudulent job types. *AI & Soc* 38, 1013–1024 (2023). https://doi.org/10.1007/s00146-022-01469-0.

[6] Alapati, Y. K., & Sindhu, K. (2016). Combining clustering with classification: a technique to improve classification accuracy. *Lung Cancer*, 32(57), 3.

[7] Piernik, M., & Morzy, T. (2021). A study on using data clustering for feature extraction to improve the quality of classification. *Knowledge and Information Systems*, 63(6), 1771–1805. https://doi.org/10.1007/s10115-021-01572-6

[8] Vidros, S., Kolias, C., & Kambourakis, G. (2016). Online recruitment services: another playground for fraudsters. *Journal of Internet Services and Applications*, 16(1), 1-6. https://doi.org/10.1016/S1361-3723(16)30025-2

[9] Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet*, 9(1), 6. https://doi.org/10.3390/fi9010006

[10] Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security*, 10(3), 155-176. https://doi.org/10.4236/jis.2019.103009

[11] Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). ORFDetector: Ensemble learning based online recruitment fraud detection. *Proceedings of the Twelfth International Conference on Contemporary Computing (IC3)* (pp. 1-5). https://doi.org/10.1109/IC3.2019.8844879

[12] Dutta, S., & Bandyopadhyay, S. K. (2020). Fake job recruitment detection using machine learning approach. *International Journal of Engineering Trends and Technology (IJETT)*, 68(4), 48. https://doi.org/10.14445/22315381/ijett-v68i4p209s

[13] Mehboob, A., & Malik, M. S. I. (2021). Smart fraud detection framework for job recruitments. *Arabian Journal for Science and Engineering*, 46(4), 3067–3078. https://doi.org/10.1007/s13369-020-04998-2