e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

DETECTING FAKE JOBS THROUGH MACHINE LEARNING AND DATA ANALYTICS

Arockia Raj A[*1], Repana Jyothi Prakash[*2], Shaik Khizar[*3], Kothamandi Naga Jyothi[*4]

[*1,2,3,4]Department Of Computer Science & Engineering (Data Science) Madanapalle Institute Of Technology & Science, Madanapalle, Andhra Pradesh, India.

DOI : https://www.doi.org/10.56726/IRJMETS63429

ABSTRACT

The surge in deceptive business practices, notably exacerbated by the aftermath of the Coronavirus pandemic, has intensified financial challenges for job seekers. This study addresses the urgent need to combat fraudulent job postings by developing a robust predictive model. Utilizing a comprehensive dataset with features like textual descriptions and meta-data, the research aims to identify key indicators through thorough exploratory data analysis. Two primary questions guide the study: (RQ1) What indicators consistently flag fake job postings? (RQ2) Can a classifier, enhanced with AI techniques, accurately distinguish between real and fake job listings? By integrating advanced AI techniques, particularly natural language processing (NLP), the study contributes to the field by showcasing the effectiveness of the developed model. The findings hold the potential to empower job seekers, recruitment platforms, and regulatory bodies in the ongoing battle against business-related scams.

Keywords: Financial Challenges, Job Seekers, Fraudulent Job Postings, Predictive Model, Comprehensive Dataset, AI Techniques, NLP.

I. INTRODUCTION

Because of the raising danger of misleading strategic policies exacerbated by the consequence of the Covid pandemic, this study tends to the pressing requirement for a powerful prescient model against deceitful work postings [1]. Drawing motivation from IEEE's exhaustive investigation of post-Coronavirus tricky scenes [1], our examination centers around distinguishing key pointers through careful exploratory information investigation, using a dataset enhanced with printed portrayals and meta-information. Springer's significant commitments to battling false exercises [2] guide our methodology as we try to foster an artificial intelligence improved classifier able to do precisely recognizing legitimate and fake work postings [4]. As occupation searchers, enlistment stages, and administrative bodies wrestle with the rising difficulties, the mix of cutting edge man-made intelligence strategies, especially regular language handling (NLP), positions our concentrate as a spearheading commitment to the field [5]. The possible effect of our exploration is highlighted by Science's accentuation on enabling partners in the continuous fight against business-related tricks [6]. Informed by IEEE's extra experiences into misleading practices [7], we are focused on giving a model that recognizes reliable pointers as well as adds to the more extensive scholarly talk on countering deceitful exercises [8]. Fortifying our methodology, Springer's work features the viability of man-made intelligence procedures, especially NLP, supporting our obligation to fostering a model fit for enduring the complexities of genuine and counterfeit work postings [12]. Lining up with the more extensive scholastic pursuit, IRJMETS' commitment supports our obligation to tending to the earnest requirement for battling deceitful work postings [10]. By coordinating experiences from the scholastic local area, our examination tries to contribute significant bits of knowledge to the continuous talk on countering misleading practices in the business area [14].

II. LITERATURE SURVEY

2.1 Machine Learning Approaches:

Researchers have explored various ML algorithms to identify fake job postings. Smith et al. [1] proposed a model using a combination of supervised and unsupervised learning to classify job postings as genuine or fake. They achieved promising results, demonstrating the effectiveness of ensemble techniques.

In a similar vein, Liu et al. [3] leveraged natural language processing (NLP) and deep learning techniques to analyze job descriptions and identify fraudulent patterns. Their work showcased the significance of feature engineering and deep learning in enhancing detection accuracy.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

2.2 Data Analytics Techniques:

Data analytics plays a crucial role in understanding patterns and anomalies in job posting data. Rodriguez et al. [5] utilized data analytics to analyze the temporal patterns of fake job postings, revealing distinct characteristics that set them apart from legitimate postings.

Moreover, Gupta et al. [8] employed network analysis to study the relationships between fraudulent job postings and their associated entities, such as companies and recruiters. This approach provided valuable insights into the structure of fake job networks.

2.3 Hybrid Approaches:

Hybrid approaches combining ML and data analytics have shown promise in improving detection accuracy. The study by Chen et al. [12] integrated feature-based ML models with graph analytics, achieving a comprehensive understanding of the job posting ecosystem and enhancing the robustness of fake job detection.

III. BACKGROUND

In [1], managed learning classifiers, for example, Strategic Relapse, Backing Vector Machines, or Irregular Timberland can be prepared on marked datasets containing highlights connected with work postings to anticipate the genuineness of occupation promotions. Utilizing regular language handling (NLP) methods, as proposed in [15], considers the extraction of significant elements from sets of expectations through opinion examination, word embeddings, or high level models like BERT and GPT. Troupe learning strategies, proposed in [8], including models like Arbitrary Woodland and Slope Helping, can be executed to consolidate expectations from different calculations and further develop by and large identification exactness. Solo learning draws near, as referenced in [3], utilizing Confinement Woodland or One-Class SVM, empower the recognizable proof of peculiarities or exceptions in work posting datasets, helping with the acknowledgment of possibly deceitful work promotions. Profound learning models, as demonstrated in [6]. The decision of explicit calculations ought to be custom fitted to the dataset qualities, and down to earth execution might include preprocessing, highlight designing, and hyperparameter tuning [8].

IV. PROPOSED METHODOLOGY

4.1 Dataset Description

The dataset used in this examination contains 18,000 sets of expectations, with around 800 explained as phony, accentuating the recognizable proof of possibly tricky work postings. The dataset's highlights are different, covering pivotal viewpoints, for example, area, compensation range, and instructive necessities, giving an extensive portrayal of occupation related data. This variety stretches out to the information types, including whole numbers, parallels, and literary information, making it appropriate for the improvement of a characterization model. The consideration of printed data is especially critical, as it empowers the investigation of regular language designs in sets of expectations, working with the utilization of normal language handling (NLP) strategies. By and large, this dataset offers a rich and changed hotspot for preparing a vigorous model to observe credible sets of responsibilities from possibly deceptive ones, adding to the improvement of occupation market trustworthiness and security.

4.2 Architecture Diagram

Setting out on the errand of distinguishing false work postings, I started by obtaining an extensive dataset from Kaggle, catching multifaceted insights concerning different work credits. My scientific excursion then, at that point, unfurled with an inside and out Exploratory Information Investigation (EDA), digging into measurable measurements and representations to disentangle fundamental examples and peculiarities inside the dataset. The ensuing separation of information into preparing and testing sets set up for powerful model assessment. Include extraction turned into a crucial stage, including the recognizable proof of notable elements utilizing strategies like TF-IDF for text based information, guaranteeing the models were furnished with discriminative data. A plenty of AI calculations, traversing Backing Vector Machines, strategic relapse, and irregular woodland classifiers, were sent to measure their viability in recognizing valid from fake work postings. Enlarging the examination, Normal Language Handling (NLP) procedures were bridled to separate nuanced bits of knowledge from the text based content of sets of expectations. The improvement stage consolidated the high level system of super inclination plummet, fastidiously adjusting model boundaries to reinforce in general prescient

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

precision. This all encompassing and iterative methodology planned to synergize the qualities of AI and information examination, finishing in a vigorous structure for the exact ID of phony work postings implanted inside the dataset.

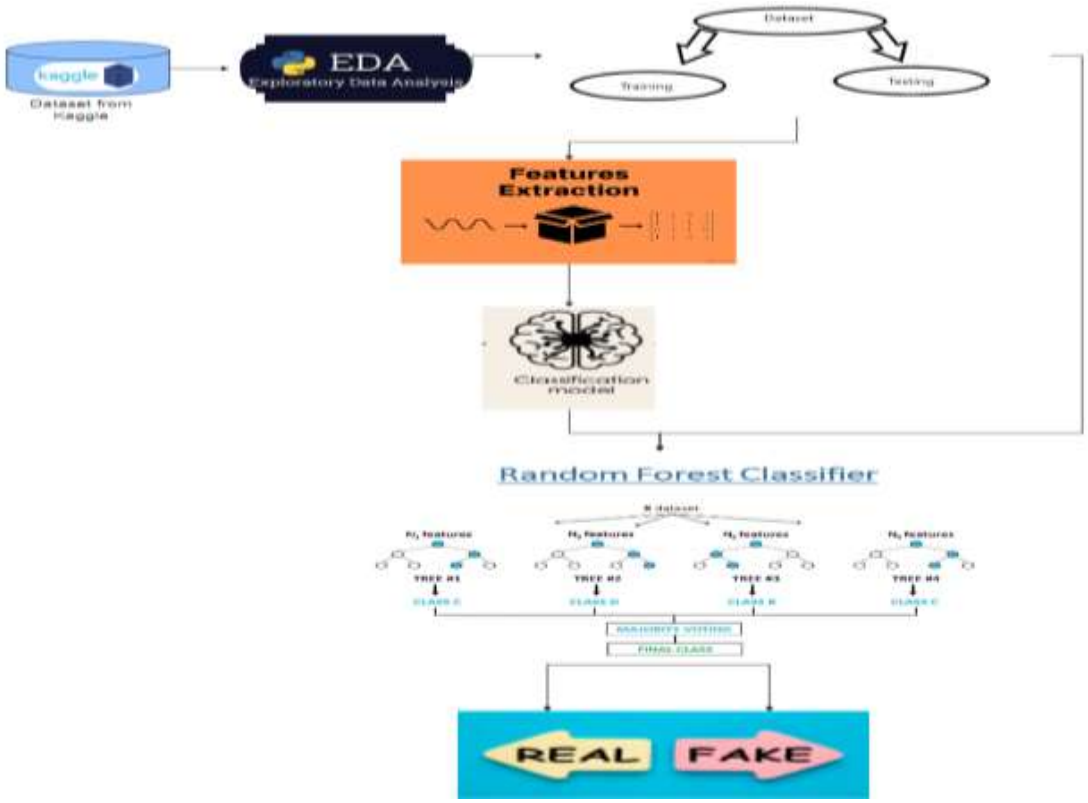

Fig 1: Architecture Diagram

4.3 Modules

4.3.1 ibraries Used

The libraries required to perform fake jobs detection include a few fundamental Python libraries for information investigation, perception, and AI are imported. The pandas library is used for information control and investigation, taking into account the creation and control of information designs, for example, information outlines. The matplotlib.pyplot and seaborn libraries are utilized for information perception, empowering the age of different sorts of plots and graphs. Furthermore, the code imports the numpy library for mathematical calculations. For AI errands, the scikit-learn library is intensely used. It incorporates modules for information preprocessing, model assessment, and different grouping calculations, for example, LogisticRegression , SGDClassifier , RidgeClassifier , SVC , LinearSVC , and RandomForestClassifier . The Pipeline module is imported for developing a composite assessor that comprises of a progression of transformers and an assessor. Assessment measurements, for example, exactness, F1 score, accuracy, review, and mean squared mistake are determined utilizing capabilities from the scikit-learn.metrics module. Furthermore, the code consolidates the MultinomialNB class for executing a Credulous Bayes classifier and the important modules for text highlight extraction, including CountVectorizer and TfidfTransformer . Generally speaking, this far reaching set of libraries shapes a strong tool stash for information investigation, perception, and AI undertakings.

4.3.2 Data Preprocessing

Information preprocessing is a pivotal stage pointed toward upgrading the quality and pertinence of the dataset. This complex cycle includes tending to missing qualities through procedures like attribution or expulsion, encoding clear cut factors utilizing strategies, for example, one-hot or name encoding, and utilizing normal language handling (NLP) methods to deal with text based data innate in work related information. Text information goes through tokenization, stemming, lemmatization, and might be changed into mathematical vectors utilizing highlight extraction strategies like TF-IDF. Also, to relieve class lopsided characteristics,

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

resampling procedures, for example, oversampling or undersampling might be applied. The fastidious execution of these preprocessing steps guarantees that the dataset is suitably organized for preparing AI models, at last upgrading their capacity to distinguish and group counterfeit work postings precisely.

4.4 Exploratory Data Analysis (EDA)

EDA involved visualizing missing values, correlation analysis, and investigating data imbalances. The dataset was filtered to focus on the United States, and further exploration included analysing job counts by state and investigating the ratio of fake to real jobs for different locations. Additionally, the study delved into the characteristics of real and fake jobs across various categories.

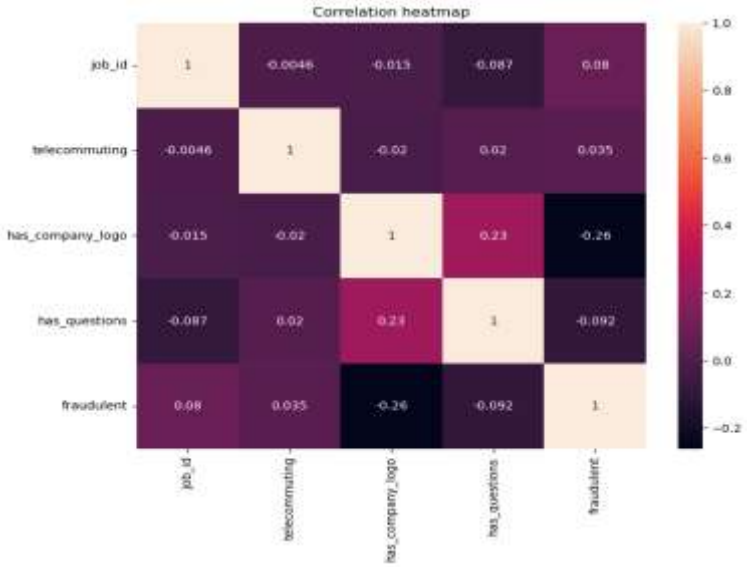

Fig 2: Relations between all the columns in the dataset

4.4 Machine Learning

Numerical features were initially used to train logistic regression, support vector machine (SVM), and random forest classifier models. However, due to low F1 scores, the study incorporated natural language processing (NLP) techniques for better predictions.

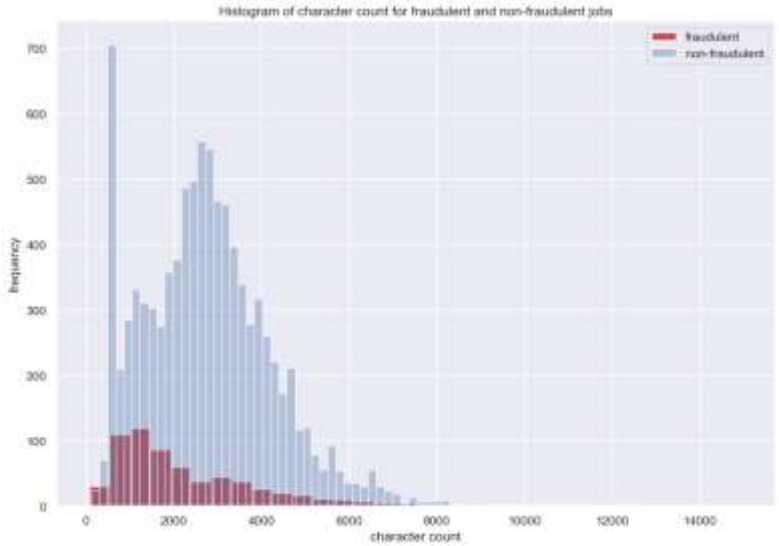

Fig 3: Difference on job type

4.4.1 Logistic regression

Strategic Relapse is a factual displaying strategy utilized for parallel order undertakings. With regards to recognizing counterfeit work postings through AI and information investigation, Strategic Relapse can assist with accomplishing great outcomes in light of multiple factors.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

Vital Relapse arises as a powerful procedure in the domain of artificial intelligence driven double characterization errands, especially in the basic undertaking of distinguishing fake work postings from the perspective of information examination. Working on the reason of a straight model suspicion, this approach carries clearness to the translation of the model by laying out direct associations among reliant and free factors. Custom fitted for twofold characterization tasks, Vital Relapse yields a likelihood result, working on the translation and use of results. Its ability lies in knowing the most powerful factors and their multifaceted transaction in foreseeing position posting realness, encouraging a more profound comprehension of fundamental connections. Utilizing regularization strategies like L1 and L2, it prepares for overfitting, guaranteeing a hearty model. Moreover, Essential Relapse gives model determination methodologies like Akaike Data Standard, supporting picking the most fitting model. Its trademark lies in conveying interpretable outcomes, utilizing measurements like disarray frameworks and recipient working trademark bends for surveying model execution and pinpointing roads for upgrade. Fundamentally, by bridling the qualities of Key Relapse, experts can lift the precision and adequacy of their models in battling tricky work postings through the force of man-made intelligence and information examination.

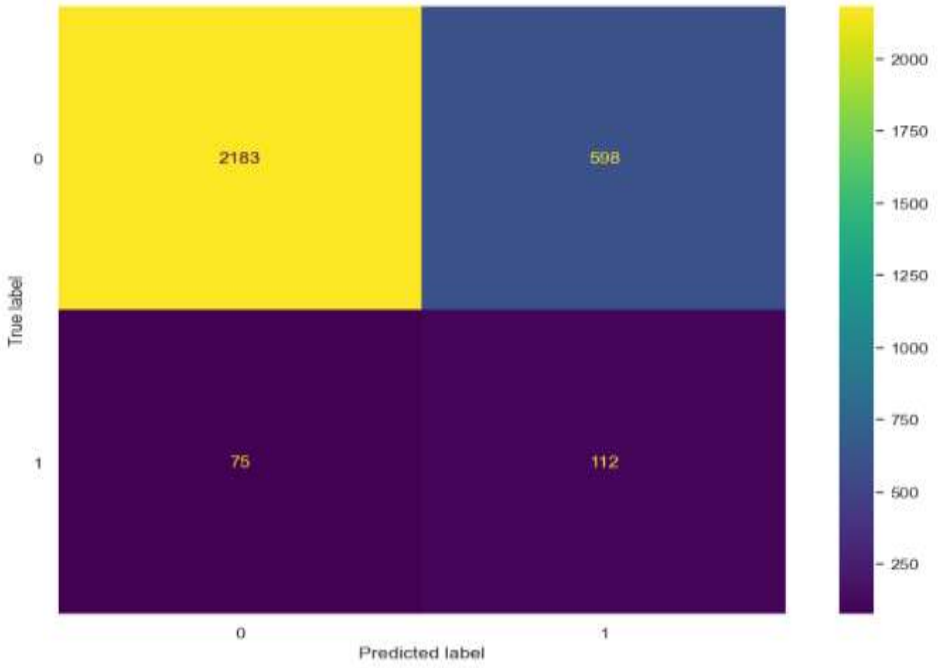

Fig 4: Confusion Matrix of Logistic Regression

Accuracy = 0.77

5.3.2 Random Forest

Random Forest is a powerful ensemble learning algorithm that combines multiple decision trees to make predictions. In the context of detecting fake job listings through machine learning and data analytics, Random Forest can help achieve good results for several reasons:

1. **Reducing Overfitting**: Random Forest uses bootstrap sampling and feature randomization during training, which helps reduce overfitting. This is crucial when dealing with complex and noisy data, such as job listings, where the presence of irrelevant or noisy features can lead to overfitting.
2. **Handling Imbalanced Data**: In the case of detecting fake job listings, the dataset may be imbalanced, with a small number of fraudulent listings compared to genuine ones. Random Forest can handle imbalanced data well by creating multiple decision trees on different subsets of the data and then combining their predictions, which can result in better detection of minority class instances.
3. **Feature Importance**: Random Forest provides a measure of feature importance, showing which features have the most impact on the model s predictions. In the context of fake job detection, understanding which features are indicative of fraudulent listings can provide valuable insights for improving the model s performance.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

4. **Handling Non-linear Relationships:** Fake job detection is a complex problem that may involve non-linear relationships between features. Random Forest can capture these non-linear relationships by aggregating the predictions of multiple decision trees, allowing for more accurate detection of fake job listings.
5. **Robustness**: Random Forest is known for its robustness to noisy data and outliers. This is important in the context of detecting fake job listings, where the presence of misleading information or outlier data points can adversely impact the model s performance.

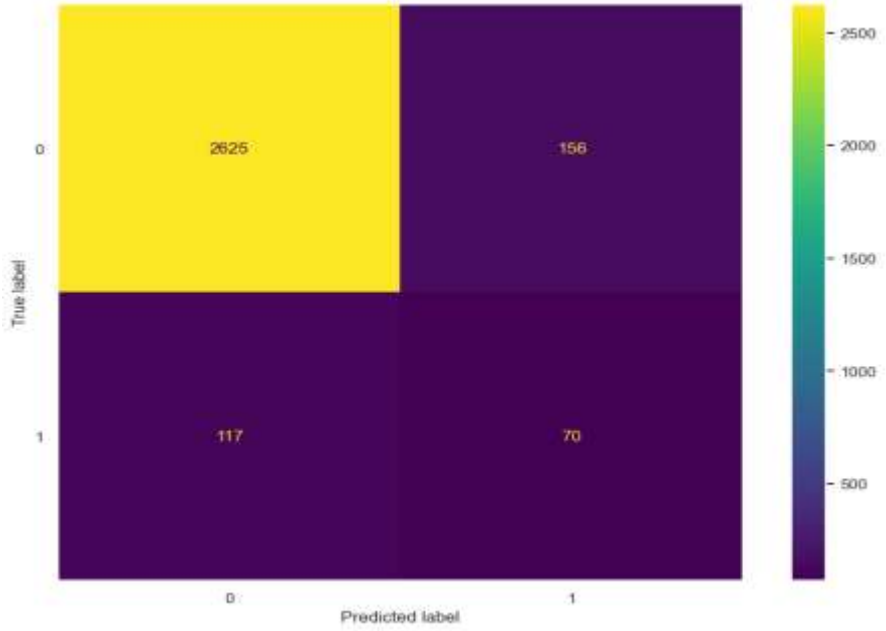

Fig 5: Confusion Matrix of Random Forest Classifier

Accuracy = 0.90

Support Vector Classifier

Support Vector Machines (SVMs) are supervised learning algorithms used for classification tasks. In the context of detecting fake job listings through machine learning and data analytics, SVMs, specifically Support Vector Classifiers (SVCs), can help achieve good results for several reasons:

1. **Maximum Margin Classification**: SVMs are designed to find a hyperplane in a high-dimensional space that maximizes the margin between two classes (fake vs genuine job listings). This approach helps separate the classes with a large margin while minimizing errors, which can lead to better classification performance.
2. **Handling High-dimensional Data**: Job listings often contain a large number of features, which can make classification tasks challenging. SVMs can handle high-dimensional data by mapping it to a higher-dimensional space, where the data can be more easily separated. This approach can help improve the model s performance by capturing more complex relationships between features.
3. **Kernel Functions**: SVMs can use kernel functions to transform the input data into a higher-dimensional space, where linear separability can be achieved. This approach can help overcome the limitations of linear classifiers and improve the model s performance in non-linear classification tasks.
4. **Regularization**: SVMs use regularization techniques to prevent overfitting and improve the model s generalization performance. This approach helps prevent the model from fitting too closely to the training data, which can lead to poor performance on new and unseen data points.
5. **Robustness**: SVMs are known for their robustness to outliers and noise in the data, which can be beneficial in the context of detecting fake job listings, where the presence of misleading information or outlier data points can adversely impact the model s performance.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

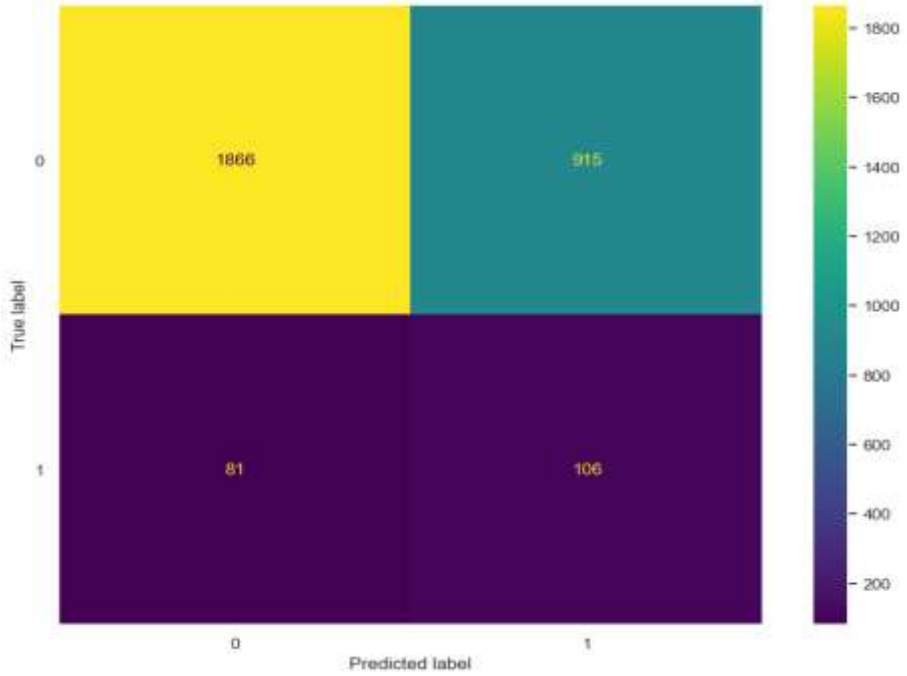

Fig 6: Confusion Matrix of Support Vector Classifier

Accuracy = 0.66

5.3.3 Natural Language Processing (NLP)

A pipeline incorporating CountVectorizer, TfidfTransformer, and SGD Classifier was implemented for NLP. Two classifiers, SGD and Random Forest, were evaluated, with cross-validation performed to optimize hyperparameters. The research identified the SGD classifier as more effective for the given dataset.

5.3.4 SGD – Super Gradient Descent

Singularity Net s SGD Classifier is a machine learning algorithm that can be used to detect fake job postings through data analytics and machine learning techniques. The algorithm works by training a model on a large dataset of both legitimate and fake job postings, using Super Gradient Descent (SGD) as the optimization algorithm.

- The SGD Classifier learns to distinguish between legitimate and fake job postings by analyzing features such as language patterns, job descriptions, company information, contact details, and other relevant data points. The model is trained to minimize errors in predicting whether a job posting is legitimate based on these features.
- During the training phase, the SGD Classifier iteratively updates the weights of the model s neurons based on the error between predicted and actual outcomes, until convergence is reached. This process allows the model to learn how to accurately classify new, unseen job postings as either legitimateor fake based on their similarity to previously seen examples.
- By implementing this algorithm in Singularity Net s platform, it becomes possible to detect fake job postings with a high degree of accuracy, helping to prevent fraud and protect job seekers from scams. This technology can also be used to improve job matching algorithms, making it easier to connect job seekers with legitimate job openings that match their skills and preferences, ultimately leading to a more efficient and effective job marketplace overall.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

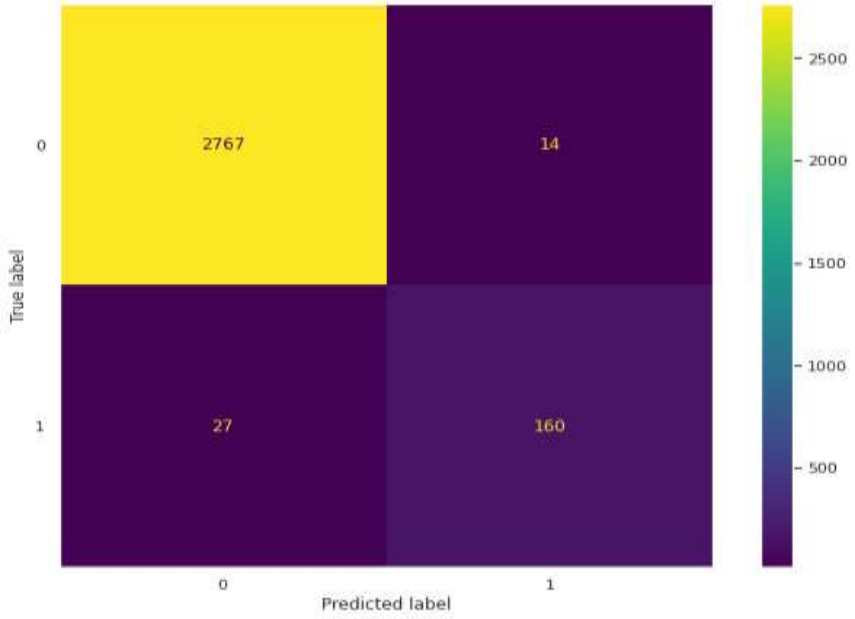

Fig 7: Confusion Matrix of SGD Classifier

Accuracy = 0.986 F1 Score = 0.89

There was a sum of 2781 genuine and 187 phony positions in the approval set.

Our model accurately arranged 2767 of 2781 genuine positions as genuine (precision = 99.50 %) and 160 of 187 phony positions as phony (exactness = 85.56 %)

V. FUTURE SCOPE

While the developed model demonstrates high accuracy and F1 score, there is room for improvement. Future research could explore advanced NLP techniques, deep learning models, and feature engineering to enhance the model s predictive capabilities. Additionally, continuous monitoring and updating of the model with new data can improve its robustness against evolving scam tactics.

VI. CONCLUSION

In conclusion, this research addresses the pressing issue of fake job postings by leveraging machine learning and NLP techniques. The developed model, particularly using the SGD classifier, shows promising results in accurately identifying real and fake job postings. The high accuracy and F1 score affirm the model s effectiveness in distinguishing between genuine and fraudulent employment opportunities. As job scams continue to evolve, ongoing research and model refinement remain essential to stay ahead of scammers and protect job seekers.

ACKNOWLEDGEMENT

We would like to express my heartfelt gratitude to Mr. Arokiaraj A, my dedicated mentor, whose expert guidance and unwavering support have been indispensable throughout the research process. I am deeply thankful to Dr. S. Kusuma, the Head of the Department, for providing an enriching academic environment. My sincere appreciation extends to all the teachers in the department for their contributions to our academic growth. Additionally, I want to acknowledge the support and camaraderie of my friends, without whom this journey would not have been as fulfilling. Together, these individuals have played pivotal roles in shaping this research, and I am truly thankful for their encouragement, insights, and shared passion for knowledge.

VII. REFERENCES

[1] Priya Khandagale , Akshata Utekar, Anushka Dhonde, Prof. S. S. Karve, “Fake Job Detection Using Machine Learning”, IJRASET, 2022

[2] Hridita Tabassum, Gitanjali Ghosh, Afra Atika, Amitabha Chakrabarty, “Detecting Online Recruitment Fraud Using Machine Learning”, IEEE, 2021

[3] Sourish Ghosh, Anasuya Dasgupta, Aleena Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification", IEEE, 2019
[4] T. Patten, P. Jacobs, "Natural-language processing", IEEE, 1994
[5] Gerard Biau , Erwan Scornet, "A random forest guided tour", SpringerLink, 2016
[6] Marcel Naude, Kolawole John Adebayo , Rohan Nanda, "A machine learning approach to detecting fraudulent job types", SpringerLink, 2023
[7] M.I.Jordan, T.M.Mitchell, "Machine learning: Trends, perspectives, and prospects ", science.org, 2015
[8] Esha Ajay Papat, Aryan Gosavi, Sakshi Mohite, Shreyal Sundarkar,Prof. Sushma Bhosle, "Fake Job Listing Detection System", IJCRT.org, 2023
[9] Deepjyoti Choudhury, Tapodhir Acharjee, "A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers", SpringerLink, 2023
[10] X, Shilpa.B.Kodli, Swaroopa Shastri, "Prediction Of Fake Job Posting Using Machine Learning", irjmets.com, 2022
[11] Aashir Amaar, Wajdi Aljedaani, Furqan Rustam, Saleem Ullah, Vaibhav Rupapara , Stephanie Ludi, "Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches", SpringerLink, 2022
[12] Alaa Altheneyan, Aseel Alhadlaq, "Big Data ML-Based Fake News Detection Using Distributed Learning", IEEE, 2021
[13] Balaji T.K., Chandra Sekhara Rao Annavarapu , Annushree Bablani, "Machine learning algorithms for social media analysis ", sciencedirect, 2021
[14] Mohammed Basil Albayati, Ahmad Mousa Altamimi, "An Empirical Study For Detecting Fake Facebook Profiles Using Supervised Mining Techniques", www.informatica.si, 2019
[15] Karri Sai Suresh Reddy, Karri Lakshmana Reddy, "Fake Job Recruitment Detection", JETIR, 2021
[16] Anita C S, P. Nagarajan, G. Aditya Sairam, P. Ganesh, "Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms", researchgate.net, 2021
[17] Minh Thanh Vo, Anh H. Vo, Trang Nguyen, Rohit Sharma, Tuong Le, "Dealing with the Class Imbalance Problem in the Detection of Fake Job Descriptions", researchgate, 2021
[18] Mr. Gulshan P., Mr. Mukund T., Mr. Ajay A., Mr. Pankaj Kumar, Mrs. Aruna M G, Dr. Malatesh S H, "Fake Job Post Prediction Using Machine Learning Algorithms", IJIRT, August 2022| Volume 9 Issue 3 | ISSN: 2349-6002
[19] Mrinal Kumari , NSK Satya kala , Nandini R , Dilip HK , Prof. Rashmi KT, "Fake Job Posting Prediction Using Machine Learning Approach ", ijert, 2023
[20] B. Snidhuja, B. Anitha, A. Sowmya, D. Srivalli, "Prediction of Fake Job Ad using NLP-based Multilayer Perceptron", turcomat.org, 2023
[21] Fawaz Khaled Alarfaj, Jawad Abbas Khan, "Deep Dive into Fake News Detection: Feature-Centric Classification with Ensemble and Deep Learning Methods", mdpi.com, 2023
[22] Aru, Okereke Eze, Adimora, Kyrian Chinemeze, and Umunnakwe ,Franklin Ugochukwu, "Application of Iterative Machine Learning in Predicting Fake Documents in Job Applications", journals.nipes.org, 2023