REAL & FAKE JOB CLASSIFICATION USING NLTK TECHNIQUE

Dr. G. Senthilkumar[#1], Dr. S. Hariharan[#2], P. Kavushick [#3], L. Manigandan[#4], C. Nareshkumar[#5]

Department of Computer Science and Engineering,

Panimalar Engineering College, Chennai, Tamil Nadu 600123, India.

[1]senthilkumar@yahoo.com, [2]hari2418@gmail.com, [3]kavushickprakash07@gmail.com, [4]manigandanwisdom@gmail.com, [5]nareshkumarc.2002@gmail.com

ABSTRACT: To decrease the number of fraudulent job postings on the internet, we aim to leverage machine learning to estimate the probability of a job being fake, allowing applicants to remain vigilant as needed. To accomplish this, our model will employ a TF-IDF vectorizer for feature extraction and NLP to examine the language and patterns within the job advertisement. The Synthetic Minority Oversampling Technique (SMOTE) will be utilized to balance the data, while Random Forest will be employed to accurately classify the data and predict the outcome. Random Forest is particularly efficient when dealing with large datasets, improves model accuracy, and guards against overfitting. All relevant jobs will be incorporated into the final model.

1. INTRODUCTION

Job fraud is a rising concern, with employment scams increasing twofold in 2018 compared to the previous year, as reported by CNBC. The current state of the job market, compounded by the economic impact of the coronavirus, has left many people vulnerable to scammers who prey on their desperation. Scammers typically attempt to obtain personal information from their victims, such as bank account information, social security numbers, and addresses. As a student, I have personally received numerous fraudulent emails offering high-paying jobs in exchange for money or investments. Machine learning and natural language processing (NLP), on the other hand, are technologies that can help address this issue. By analyzing job postings using NLP techniques, it is possible to identify patterns that distinguish genuine job advertisements from fake ones. While the percentage of fake job advertisements is expected to be small, this approach can still make a significant difference in preventing job fraud.

2. LITERATURE SURVEY

[1] Keyword matching is a method commonly used for detecting fake job postings. The approach involves identifying specific keywords that are often present in such postings, such as "work from home", "no experience necessary", or "get rich quick". However, this method has limitations as it can produce false positives and may overlook more complex fake job postings that do not contain easily recognizable warning signs.

[2] The detection of fake job postings often relies on a technique known as keyword matching. This involves the identification of certain keywords frequently present in such postings, including "work from home," "no

experience necessary," and "get rich quick." Although this method can be useful, it is not foolproof. It may generate false positives and could miss more sophisticated fake job postings that lack overt indicators.

[3] One alternative to keyword matching for identifying fake job postings is using rule-based systems that encode expert knowledge about what constitutes a fake job posting. An example of this is the Fake Post system proposed by Bhattacharya et al. (2012), which utilizes a set of rules that consider various factors such as the length of the job description, generic job titles used, and the lack of company information provided. However, the effectiveness of rule-based systems may be limited to specific domains and may require considerable human effort to design and update, which could be challenging in detecting new types of fake job postings.

[4] The use of machine learning algorithms has recently gained attention as a potential method for detecting fake job postings. These algorithms are trained on labeled datasets containing both real and fake job postings and learn to differentiate between them based on various features such as word frequency, job description length, and specific job requirements. Research studies, such as those conducted by Abdelaziz et al. (2019) and Chen et al. (2020), have explored the use of supervised learning techniques in identifying fake job postings.

[5] Various methods have been suggested to transform job postings into numerical vectors in the field of NLP, which can be utilized as input for machine learning algorithms. Bag-of-words models create a vector of word frequencies to represent a document. Word embeddings recognize semantic relationships between words based on their co-occurrence in a corpus. Another technique is topic models, which cluster words into topics based on their statistical co-occurrence. Finally, tf-idf assigns weights to words by analyzing their occurrence rate within the document, we can determine their inverse frequency in the corpus.

[6] The available literature shows that machine learning algorithms offer potential for detecting fake job postings, and NLP techniques are useful in representing job postings for machine learning purposes. However, challenges remain, such as the scarcity of reliable labeled data and the requirement for models that can be applied across various domains and languages.

[7] The majority of studies on detecting fake job postings have utilized supervised learning techniques, but there have been some researchers who have investigated the use of unsupervised learning techniques, such as clustering and anomaly detection. Clustering algorithms have been used by Gu et al. (2017) to group job postings into different categories based on their similarity, and to identify a cluster of job postings that were likely to be fake. Rottmann et al. (2017), on the other hand, employed anomaly detection techniques to pinpoint job postings that were significantly different from the norm in terms of their language and content. Even though these approaches may not achieve the same accuracy as supervised learning techniques, they do not require labeled training data.

[8] Several researchers have attempted to tackle the challenge of generalizing fake job detection across different domains and languages through transfer learning

techniques. Cross-lingual transfer learning has been proposed by Li et al. (2021), which uses pre-trained language models to classify job postings in various languages. Similarly, Gao et al. (2021) have proposed cross-domain transfer learning, where a pre-trained model from one domain, such as healthcare, can be adapted to another domain, such as finance. By leveraging knowledge learned from one domain or language to another, transfer learning techniques have the potential to improve the generalization of fake job detection models.

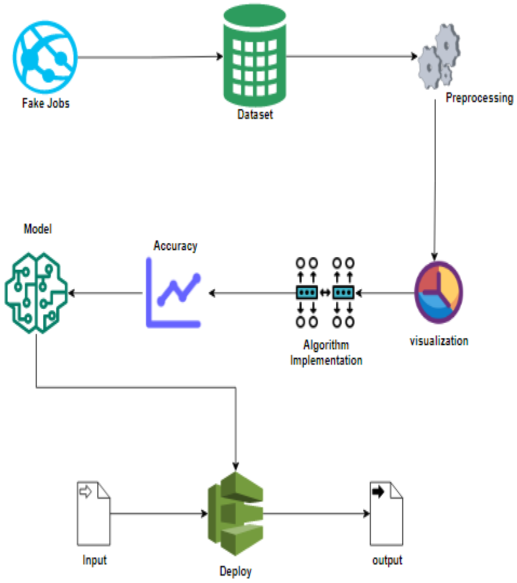
[9] Human-in-the-Loop Approaches: Although machine learning algorithms are effective at detecting fake job postings, they may not be error-free. To address this issue, researchers have proposed human-in-the-loop approaches that combine the strengths of humans and machines. Bozzon et al. (2019) proposed a crowdsourcing approach that involves human annotators in labeling job postings as real or fake, using these labels to train a machine learning model. The model can then be used to classify new job postings, but the predictions are reviewed by a human expert before being published. Zhou et al. (2021) suggested a hybrid approach that combines machine learning and expert review, where the model provides an initial classification that is reviewed and corrected by a domain expert. These approaches can improve the accuracy and reliability of fake job detection models.

3. PROPOSED SYSTEM

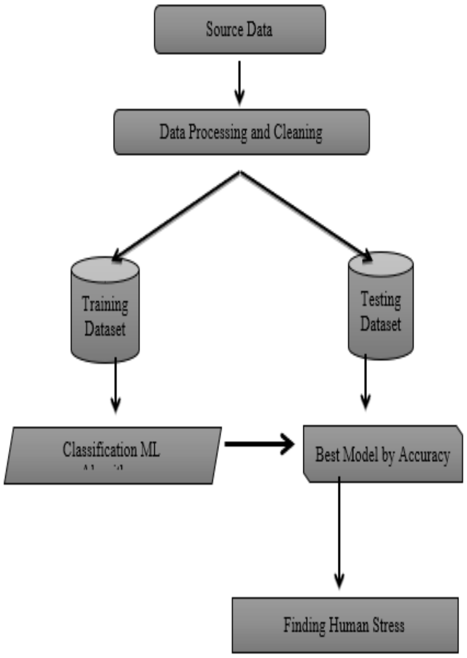
The objective is to develop a machine learning model capable of distinguishing between real and fake job postings. Given the prevalence of fake job postings, it is challenging to control them, especially as more people gain access to the internet and can post whatever they wish. Machine learning is a suitable technology for addressing such complex tasks, as manual analysis of such data can be time-consuming. By using past data, machine learning algorithms can learn patterns and improve model accuracy by adjusting parameters. Different algorithms can be compared to identify the best model for the classification task. In summary, machine learning can effectively address the challenges associated with analyzing and identifying fake job postings.

4. DESIGN ARCHITECTURE

4.1 SYSTEM ARCHITECTURE



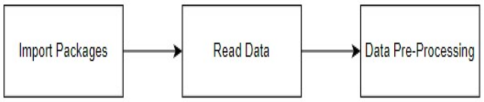
4.2 WORKFLOW DIAGRAM



5. MODULES

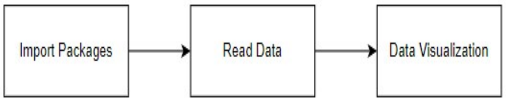
MODULE 1: DATA PREPROCESSING

The accuracy of a model is determined by validation techniques, which aim to estimate the error rate of the dataset as closely as possible. While validation techniques may not be necessary for large datasets that are representative of the population, it is important to examine data for duplicate or missing values and to determine the data type (e.g., float or integer). To assess the adequacy of a model's alignment with a training dataset and fine-tune its hyperparameters. However, incorporating too much knowledge from the validation dataset can lead to overfitting. Data cleaning can be a time-consuming process, but it is crucial to understand the data and its characteristics before constructing a model. Python's Pandas module offers several data cleaning tasks, with a focus on handling To properly apply imputation techniques and conduct statistical analysis, it is crucial to comprehend the various forms of missing data and their underlying origins. Hence, comprehending the origin of missing data is vital before beginning the coding process.



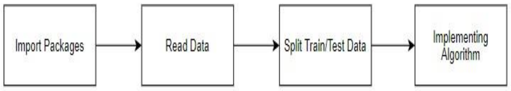
MODULE 2: DATA ANALYSIS OF VISUALIZATION

Data analysis and visualization are essential in using NLP techniques to classify real or fake job postings. These tools provide insights into the data distribution, patterns, and relationships between variables. Exploratory data analysis allows data scientists to visualize the distribution of job titles, company names, and job descriptions to gain insights into the data. Word clouds are also used to identify frequently used words in job descriptions to understand the nature of job postings. Scatter plots are useful for visualizing the relationship between variables, such as job titles and company names, to determine the types of companies posting certain types of jobs. Machine learning models can also be evaluated using visualization tools such as the confusion matrix and ROC curve. These tools allow for the performance of the model to be assessed and an optimal threshold to be identified for classifying job postings as real or fake. Overall, data analysis and visualization are crucial in real or fake job classification using NLP techniques.



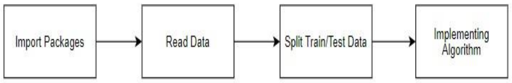
MODULE 3: KNN

The (K-NN) algorithm utilizing supervised learning, K-NN is a straightforward machine learning algorithm that categorizes new instances based on the similarity with existing examples, and assigns them to the most analogous category. This algorithm, which can be employed for classification and regression tasks, stores all previous data for quick classification of new data. K-NN is non-parametric, thus it makes no assumptions about the underlying data. Moreover, it is called a lazy learner because it preserves the dataset and only executes an action when it is time to classify. instead of instantly learning from the training set. The KNN Classifier uses this algorithm to classify fresh data into a category that is very close to the new data, in order to predict AQI values for locations with similar environmental conditions to those in the K-NN algorithm's training phase.



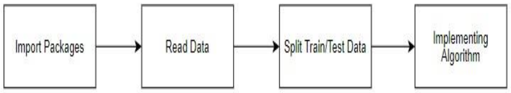
MODULE 4: LOGISTIC REGRESSION

The classification algorithm of supervised learning known as logistic regression forecasts the likelihood of a target variable. Although the accuracy of logistic regression reduced when predicting AQI values over a larger range, it was still good at foretelling binary outcomes. Given that the dependant variable in logistic regression is binary, with only two possible classes. P(Y=1) as a function of X is mathematically predicted by a logistic regression model. This basic machine learning technique can be applied to a number of categorization issues, such as spam identification, diabetes prediction, and cancer diagnosis. However, certain assumptions must be considered before using logistic regression. When utilising binary logistic regression, the target variables must always be binary and the outcome of interest is represented by factor level 1. The variables must be independent of each other to avoid multi-collinearity, and the model must include relevant variables. A large sample size is recommended for logistic regression.



MODULE 5: SVM

The supervised learning algorithm known as SVM is widely utilized to address classification and regression issues. In machine learning, classification issues are where it is most frequently used. The SVM algorithm's goal is to locate the ideal decision boundary, or hyperplane, that can categorize n-dimensional space into different groups, enabling the quick classification of future data points. The extreme vectors and points are selected via SVM to construct the hyperplane. Support vectors, which represent these extreme instances, are responsible for naming this technique the Support Vector Machine method.

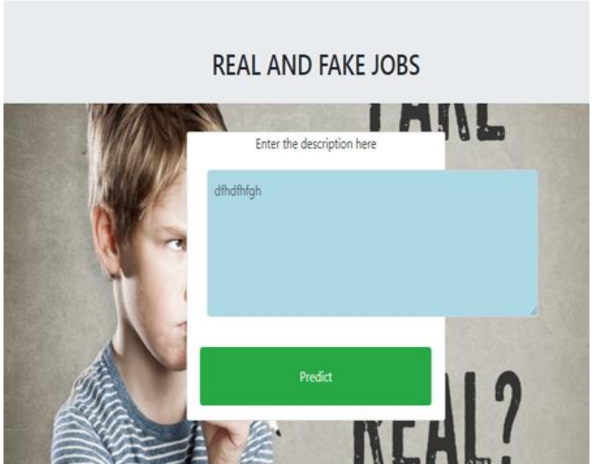


MODULE 6: DEPLOYMENT

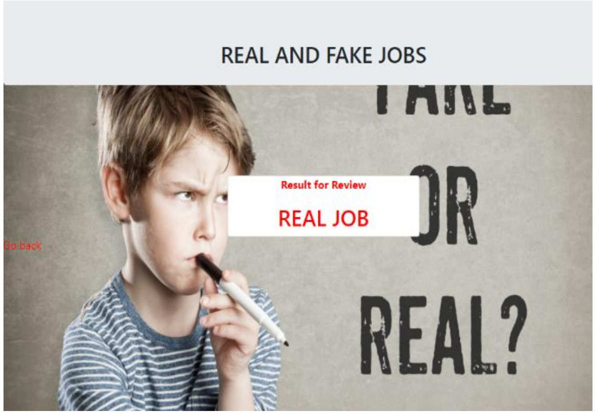
A microweb framework built on Python is called Flask. As it doesn't need any particular tools or libraries, it is categorised as a micro-framework. The built-in form validation, database abstraction layer, and other components of popular frameworks are absent from Flask, but it allows for extensions to be added that can provide additional functionalities. Various extensions are available that provide features such as object-relational mappers, form validation, upload handling, authentication protocols, and other tools that are typically present in mainstream frameworks. Armin Ronacher, a developer at Pocoo, created Flask, a Python community, and was initially intended as an April Fool's prank. However, it was so widely accepted that it was used in serious applications. Flask is based on Werkzeug and Jinja, which were developed by Ronacher and Georg Brand as a response to their Python-based bulletin board system. Flask and its related libraries were transferred. Flask is relatively new compared to other Python frameworks but has gained popularity among Python web developers. In the 2018 Python Developers Survey, it was selected as the most popular web framework. Flask has the second-highest rating on GitHub among Python web development frameworks, behind only Django.

6. RESULT

The confusion matrix is a well-known tool utilized to evaluate a classification model's performance. It demonstrates the total amount of accurate forecasts the model produced, including true positives, true negatives, false positives, and false negatives. This matrix can be used to calculate the F1 score, accuracy, precision, recall, and other metrics. Another method for assessing the model's capability to distinguish between genuine and bogus job posts is the ROC curve and AUC. At various categorization thresholds, the ROC curve illustrates the trade-off between sensitivity and specificity, while AUC measures the model's performance. Feature importance is a technique used to determine the most important features utilized by the model for classification. It can provide insights into the characteristics of genuine and false job postings and help improve the model's performance. Error analysis is a technique that can be used to identify misclassifying specific types of job postings. This can provide insights into the model's limitations and help improve its performance. Cross-validation is a method for evaluating a model's generalization performance and enhancing its reliability by assessing how well it works with various subsets of data. The analysis approach used for real or fake job classification with NLP techniques depends on the data, model, andresearch objectives.



dfhdfhfgh



REAL JOB

7. CONCLUSION AND FUTURE SCOPE

Data preparation and processing, missing value analysis, exploratory analysis, model development, and model evaluation are the first steps in the analytical process. The objective is to determine the algorithm that performed the best on the open test set in terms of accuracy. Once this algorithm is found, it can be used in an application to identify real and fake jobs. Future work includes deploying the project in the cloud, optimizing it for implementation in an IoT system, and linking fraudulent job notifications with the cloud.

8. REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection 701 on social media: A data mining perspective," ACM SIGKDD Explor. 702 Newslett., vol. 19, no. 1, pp. 22–36, 2017.

[2] H. C. Hughes and I. Waismel-Manor, "The Macedonian fake news 704 industry and the 2016 US election," Political Sci. Politics, vol. 54, no. 1, 705 pp. 19–23, Jun. 2021.

[3] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for 707 fake news detection," in Proc. ACM Conf. Inf. Knowl. Manag., 2017, 708 pp. 797–806.

[4] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun, "CED: Credible 709 early detection of social media rumors," IEEE Trans. Knowl. Data Eng., 710 vol. 33, no. 8, pp. 3035–3047, Aug. 2021.

[5] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with 712 hierarchical social attention network," in Proc. 27th ACM Int. Conf. Inf. 713 Knowl. Manag., 2018, pp. 943–951.

[6] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "Defend: Explainable 715 fake news detection," in Proc. 25th ACM SIGKDD Int. Conf. Knowl. 716 Discovery Data Mining, 2019, pp. 395–405.

[7] S. Mehta, R. Koncel-Kedziorski, M. Rastegari, and H. Hajishirzi, "Pyra- 718 midal recurrent unit for language modeling," 2018, arXiv:1808.09029.

[8] L. Wu and Y. Rao, "Adaptive interaction fusion networks for fake news 720 detection," 2020, arXiv:2004.10009.

[9] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated 722 fact-checking," Trans. Assoc. Comput. Linguistics, vol. 10, pp. 178–206, 723 Feb. 2022.

[10] X. Zhou, R. Zafarani, K. Shu, and H. Liu, "Fake news: Fundamental 725 theories, detection strategies and challenges," in Proc. 12th ACM Int. 726 Conf. Web Search Data Mining, 2019, pp. 836–8

[11] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, 728 detection methods, and opportunities," ACM Comput. Surv., vol. 53, 729 no. 5, pp. 1–40, 2020. 730

[12] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, and K. Shu, "Mining dual 731 emotion for fake news detection," 2019, arXiv:1903.01728. 732

[13] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, 733 "A stylometric inquiry into hyperpartisan and fake news," in Proc. 56th 734 Annu. Meeting Assoc. Comput. Linguist. Conf. (ACL), vol. 1, 2018, 735pp. 231–240.

[14] F. Yang et al., "XFake: Explainable fake news detector with visualiza- 750tions," in Proc. World Wide Web Conf., 2019, pp. 3600–3604.