ISSN 1533-9211

DOI 10.5281/zenodo.6698012

FAKE JOB RECOMMENDATION SYSTEM

[1] P. KIRUTHIKA, [2] N.P. PREETHI, [3] S. ANISHA, [4] M. MURUGESWARI and

[5] B. VIJAYALAKSHMI

[1] Assistant Professor, Department of Computer Science and Engineering, RVS Technical Campus-Coimbatore.
[2345] Students of Department of Computer Science and Engineering, RVS Technical Campus-Coimbatore.

ABSTRACT

To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts. Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers. The fake news on social media and various other media is wide spreading and is a matter of serious concern due to its ability to cause a lot of social and national damage with destructive impacts.

1. INTRODUCTION

Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF). In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the job-seekers. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. Fraudulent job advertisements can be posted against a reputed company for violating their credibility. These fraudulent job post detection draws a good attention for obtaining an automated tool for identifying fake jobs and reporting them to people for avoiding application for such jobs. For this purpose, machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques are considered initially. A classifier maps input variable to target classes by considering training data. Classifiers addressed in the Project for identifying fake job posts from the others are described briefly. These classifiers are based prediction may be broadly categorized into –Single Classifier based Prediction and Ensemble Classifiers based Prediction. And real Job Recommendation system.

2. LITERATURE SURVEY

2.1 Fake Job Recruitment Detection Using Machine Learning Approach

To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed in the paper. Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts. Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers.

2.2 A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques

In recent years, due to advancement in modern technology and social communication, advertising new job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perception and deep neural network to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post.

3. PROBLEM STATEMENTS

- This project aims to create a classifier that will have the capability to identify fake and real jobs. The final result is evaluated based on two different models. Since the data provided has numeric and features, one model will be used on the text data and another on numeric data. The final output will be a combination of the two. The final model will take in any relevant job posting data and produce a final result determining whether the job is real or not. And Real Job Recommendation

3.1 Existing System

Previous model and the methodologies, to create the ORF detection model where we have used our own dataset. We have created our dataset based on the job field and by using a publicly accessible dataset as a reference. Furthermore, Logistic Regression, AdaBoost detecting fraudulent.

3.2 Disadvantages

- The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. It not only provides a measure of how appropriate a predictor (coefficient size) is, but also its direction of association (positive or negative).
- AdaBoost is being used to classify text and images rather than binary classification problems. The main disadvantage of AdaBoost is that it needs a quality dataset. Noisy data and outliers have to be avoided before adopting an AdaBoost algorithm.

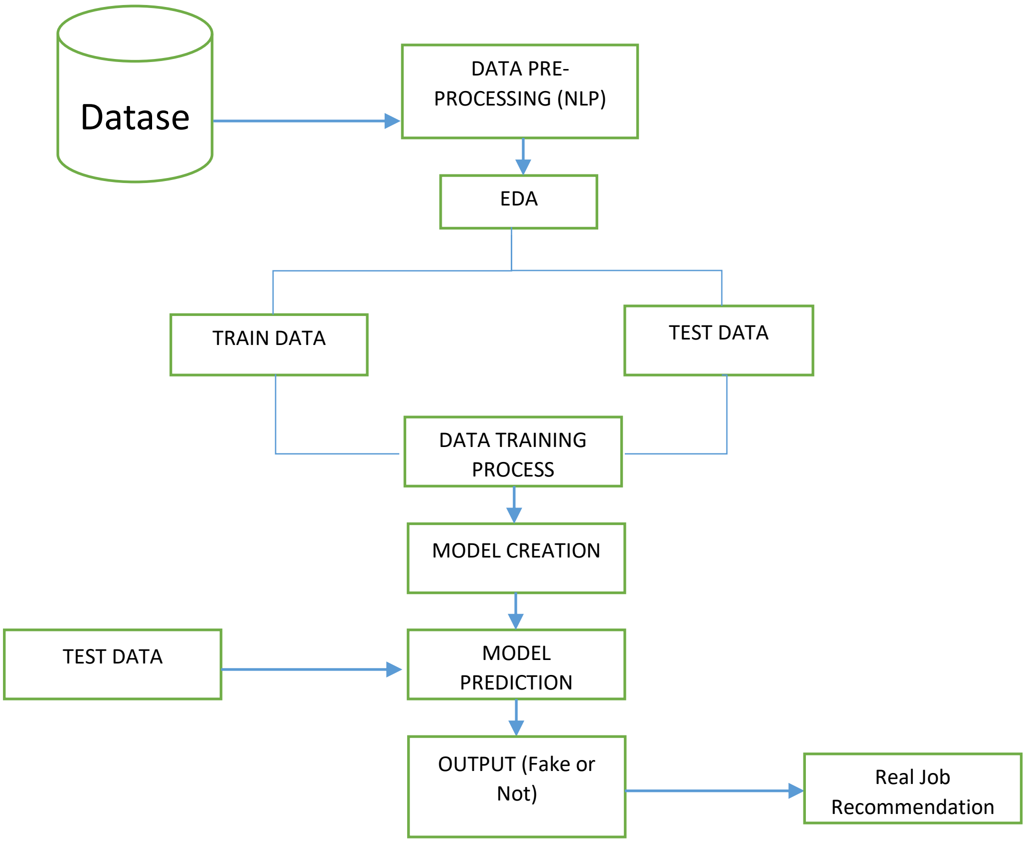
The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6698012

4. PROPODED SYSTEM

Input variable to target classes by considering training data. Classifiers addressed in the Project for identifying fake job posts from the others are described briefly. This classifiers-based prediction may be broadly categorized into – Single Classifier based Prediction and Ensemble Classifiers based Prediction and real Job Recommendation system. In this Project, we are using Random Forest classifier with Machine Learning.

- Impressive in Versatility.
- Parallelizable. They are parallelizable, meaning that we can split the process to multiple machines to run.
- Great with High dimensionality.
- Quick Prediction/Training Speed.
- Robust to Outliers and Non-linear Data.
- Handles Unbalanced Data.
- Low Bias, Moderate Variance.

4.1 Flow Diagram



4.2 Advantages

- Impressive in Versatility.
- Parallelizable. They are parallelizable, meaning that we can split the process to multiple machines to run.
- Great with High dimensionality.
- Quick Prediction/Training Speed.
- Robust to Outliers and Non-linear Data.
- Handles Unbalanced Data.
- Low Bias, Moderate Variance.

5. SOFTWARE REQUIREMENTS

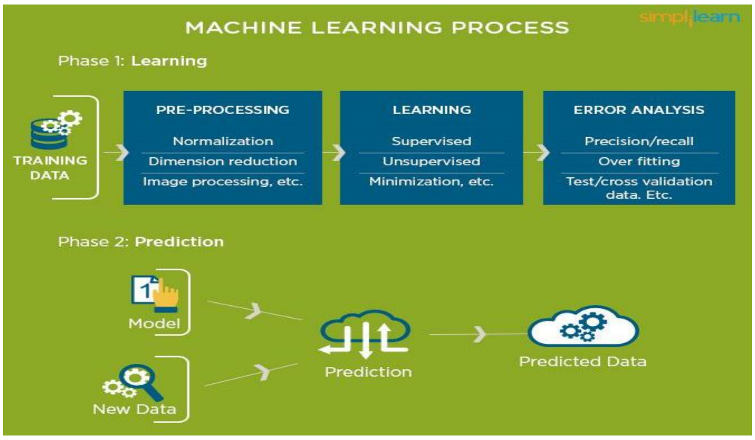
- Machine learning
- Python

5.1 Software Modules

Machine learning

Why machine learning?

To better understand the uses of machine learning consider some of the instances where machine learning is applied: the self-driving Google car, cyber fraud detection, online recommendation engines—like friend suggestions on Facebook, Netflix showcasing the movies and shows you might like, and “more items to consider” and “get yourself a little something” on Amazon—are all examples of applied machine learning.



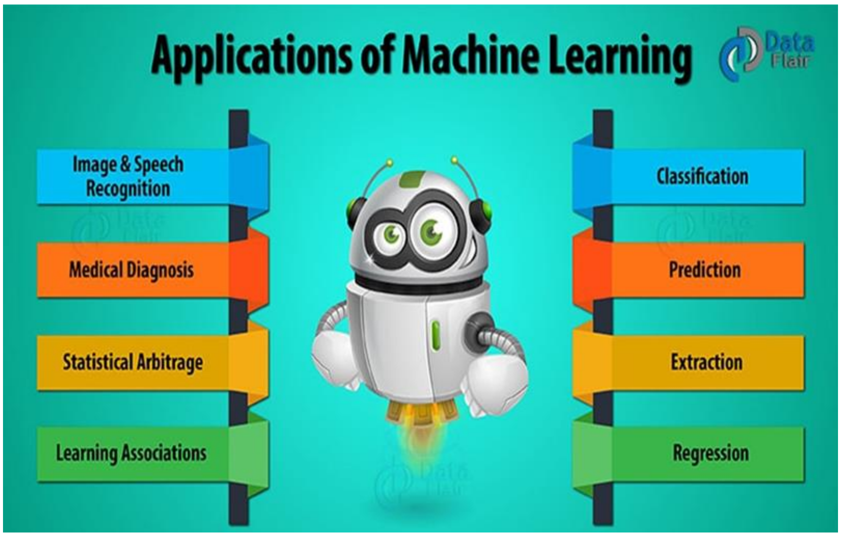
Machine learning has also changed the way data extraction, and interpretation is done by involving automatic sets of generic methods that have replaced traditional statistical techniques.

The Seybold REPORT

ISSN 1533-9211

DOI 10.5281/zenodo.6698012

Application of Machine Learning



Python

Python is an open-source programming language. Python was made to be easy-to-read and powerful. A Dutch programmer named Guido van Rossum made Python in 1991. He named it after the television show Monty Python's Flying Circus. Many Python examples and tutorials include jokes from the show. Its standard library is made up of many functions that come with Python when it is installed.

Some things that Python is often used for are:

- Web development
- Game programming
- Desktop GUIs
- Scientific programming

Python 3.0 (also called "Python 3000" or "Py3K") was released on December 3, 2008 It was designed to rectify fundamental design flaws in the language—the changes required could not be implemented while retaining full backwards compatibility with the 2.x series, which necessitated a new major version number. The guiding principle of Python 3 was: "reduce feature duplication by removing old ways of doing things".

6. METHODOLOGIES

Principal Component Analysis

- The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent.
- The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered

DOI 10.5281/zenodo.6698012

such that the retention of variation present in the original variables decreases as we move down in the order.

Classification

- In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. it is a technique where we categorize data into a given number of classes. ... Classification model: A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. For classification we use machine learning algorithm .as well as prediction.

Prediction

Machine learning is a way of identifying patterns in data and using them to automatically make predictions or decisions. The two main method of machine learning you will focus on are regression and classification. Here we predict Real Job or Fake Job, mild case etc. by algorithm performance.

Confusion Matrix

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

- The confusion matrix shows the ways in which your classification model is confused when it makes predictions.
It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

Here,

- Class 1 :Positive
- Class 2 :Negative

Definition of the Terms:

- Positive (P): Observation is positive (for example: is an apple).
- Negative (N): Observation is not positive (for example: is not an apple).
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Classification Rate/Accuracy:

Classification Rate or Accuracy is given by the relation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

Recall: Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given by the relation:

$$\textit{Recall} = \frac{TP}{TP + FN}$$

Precision: To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number of FP).

Precision is given by the relation:

$$\textit{Precision} = \frac{TP}{TP + FP}$$

High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

F-measure: Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.

The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$\textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

In which we have infinite data elements of class B and a single element of class A and the model is predicting class A against all the instances in the test data. Here,

- Precision: 0.0
- Recall: 1.0

The Seybold REPORT

ISSN 1533-9211

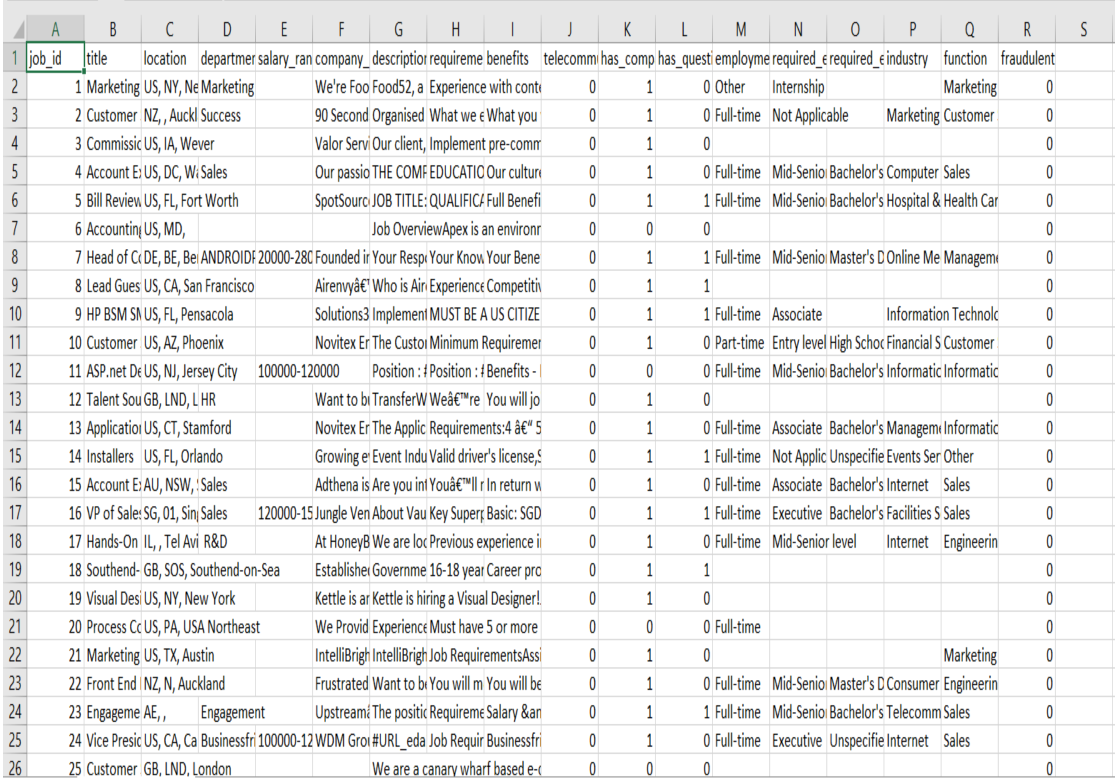
DOI 10.5281/zenodo.6698012

7. MODULES

- Data set description
- Feature selection
- Algorithm Implementation
- Classifier
- Recommendation.

Data set description

The dataset should be collected from the Kaggle website. If the given dataset of the job is true and original, it will recommend the job. If the given dataset of the job is fake, then the model will find the original job related to the fake job.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	job_id	title	location	departmer	salary_ran	company_	descriptior	requireme	benefits	telecomm	has_comp	has_questi	employme	required_e	required_e	industry	function	fraudulent	
2	1	Marketing	US, NY, Ne	Marketing		We're Foo	Food52, a	Experience with cont		0	1	0	Other	Internship			Marketing	0	
3	2	Customer	NZ, , Auckl	Success		90 Second	Organised	What we e	What you	0	1	0	Full-time	Not Applicable		Marketing	Customer	0	
4	3	Commissic	US, IA, Wever			Valor Servi	Our client,	Implement	pre-comm	0	1	0						0	
5	4	Account E	US, DC, W	Sales		Our passio	THE COMF	EDUCATIO	Our cultur	0	1	0	Full-time	Mid-Senio	Bachelor's	Computer	Sales	0	
6	5	Bill Review	US, FL, Fort Worth			SpotSourc	JOB TITLE:	QUALIFICA	Full Benefi	0	1	1	Full-time	Mid-Senio	Bachelor's	Hospital &	Health Car	0	
7	6	Accountin	US, MD,				Job OverviewApex is an environr			0	0	0						0	
8	7	Head of C	DE, BE, Be	ANDROIDF	20000-280	Founded ir	Your Resp	Your Know	Your Bene	0	1	1	Full-time	Mid-Senio	Master's D	Online Me	Managem	0	
9	8	Lead Gues	US, CA, San Francisco			Airenvyâ€	Who is Air	Experience	Competitiv	0	1	1						0	
10	9	HP BSM SN	US, FL, Pensacola			Solutions3	Implement	MUST BE A US CITIZE		0	1	1	Full-time	Associate		Information Technolc		0	
11	10	Customer	US, AZ, Phoenix			Novitex Er	The Custor	Minimum Requiremer		0	1	0	Part-time	Entry level	High Schoc	Financial S	Customer	0	
12	11	ASP.net De	US, NJ, Jersey City		100000-120000		Position :	Position :	Benefits -	0	0	0	Full-time	Mid-Senio	Bachelor's	Informatic	Informatic	0	
13	12	Talent Sou	GB, LND, L	HR		Want to b	TransferW	Weâ€™re	You will jo	0	1	0						0	
14	13	Applicatio	US, CT, Stamford			Novitex Er	The Applic	Requirements:4 â€" 5		0	1	0	Full-time	Associate	Bachelor's	Managem	Informatic	0	
15	14	Installers	US, FL, Orlando			Growing e	Event Indu	Valid driver's license,S		0	1	1	Full-time	Not Applic	Unspecifie	Events Ser	Other	0	
16	15	Account E	AU, NSW,	Sales		Adthena is	Are you int	Youâ€™ll r	In return w	0	1	0	Full-time	Associate	Bachelor's	Internet	Sales	0	
17	16	VP of Sale	SG, 01, Sin	Sales	120000-15	Jungle Ven	About Vau	Key Superp	Basic: SGD	0	1	1	Full-time	Executive	Bachelor's	Facilities S	Sales	0	
18	17	Hands-On	IL, , Tel Avi	R&D		At HoneyB	We are loc	Previous experience i		0	1	0	Full-time	Mid-Senior level		Internet	Engineerin	0	
19	18	Southend-	GB, SOS, Southend-on-Sea			Establishe	Governme	16-18 year	Career pro	0	1	1						0	
20	19	Visual Desi	US, NY, New York			Kettle is ar	Kettle is hiring a Visual Designer!.			0	1	0						0	
21	20	Process Cc	US, PA, USA Northeast			We Provid	Experience	Must have 5 or more		0	0	0	Full-time					0	
22	21	Marketing	US, TX, Austin			IntelliBrigh	IntelliBrigh	Job RequirementsAssi		0	1	0					Marketing	0	
23	22	Front End	NZ, N, Auckland			Frustrated	Want to b	You will m	You will be	0	1	0	Full-time	Mid-Senio	Master's D	Consumer	Engineerin	0	
24	23	Engageme	AE, ,	Engagement		Upstream	The positic	Requireme	Salary &an	0	1	1	Full-time	Mid-Senio	Bachelor's	Telecomm	Sales	0	
25	24	Vice Presic	US, CA, Ca	Businessfri	100000-12	WDM Grou	#URL_eda	Job Requir	Businessfri	0	1	0	Full-time	Executive	Unspecifie	Internet	Sales	0	
26	25	Customer	GB, LND, London				We are a canary wharf based e-(			0	0	0						0	

Feature Selection

- Feature selection is the process of reducing the number of input variables when developing a predictive model.

The Seybold REPORT

ISSN 1533-9211

DOI 10.5281/zenodo.6698012

Algorithm Implementation

- Implementing a machine learning algorithm will give you a deep and practical appreciation for how the algorithm works.
- There are numerous micro-decisions required when implementing a machine learning algorithm and these decisions are often missing from the formal algorithm descriptions.

Random Forests Classifiers

- Random forests is a supervised learning algorithm. It can be used both for classification and regression.
- It is also the most flexible and easy to use algorithm. A forest is comprised of trees.
- It is said that the more trees it has, the more robust a forest is Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Natural Language Processing

- The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP).
- It also includes graphical demonstrations and sample data sets as well as accompanied by a cook book and a book which explains the principles behind the underlying language processing tasks that NLTK supports.

Classifier

A classifier in machine learning is an algorithm that automatically orders or categorizes data into one or more of a set of “classes.” One of the most common examples is an email classifier that scans emails to filter them by class label: Spam or Not Spam.

Recommendation

- The content-based method is a domain-dependent algorithm which focuses on much more on the evaluation of the characteristics of things to produce predictions.
- When files like pages, publications as well as news are being suggested,

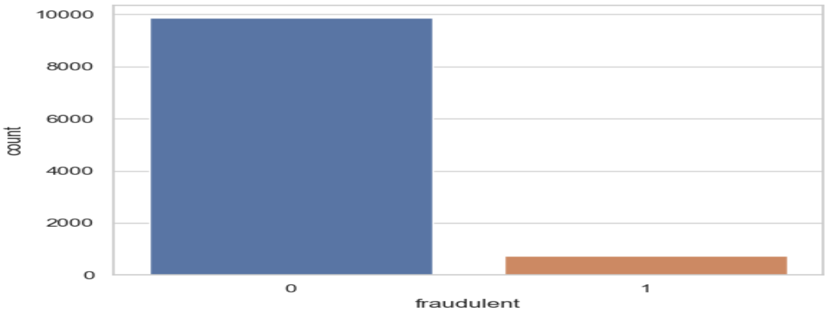
Chart Output:

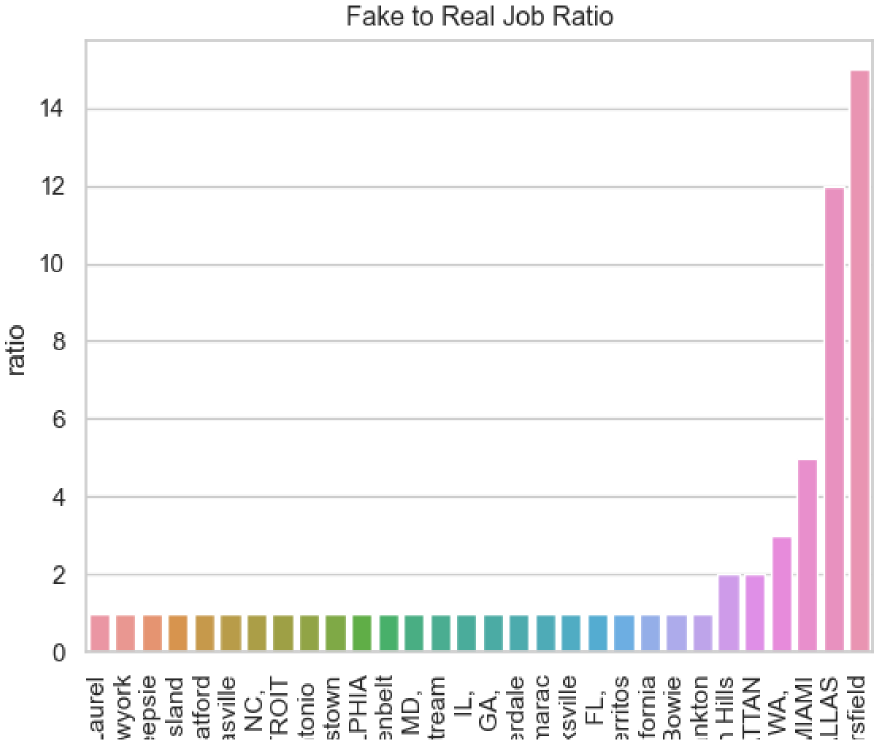
Fake job and Real Job Count:

The Seybold Report

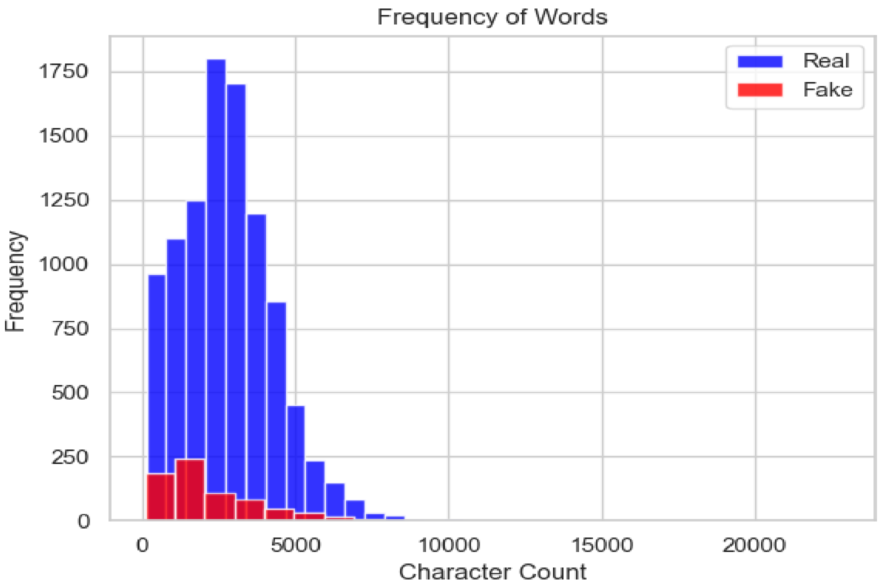
ISSN 1533-9211

DOI 10.5281/zenodo.6698012





DOI 10.5281/zenodo.6698012



8. CONCLUSIONS

Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that Random Forest classifier outperforms over its peer classification tool. The proposed approach achieved accuracy 97.27% accurate. Based on the obtained results we recommended Real Job for different level of Domain.

Reference

- Gulshan Shrivastava, Member, IEEE, Prabhat Kumar, Senior Member, IEEE, Rudra Pratap, Pramod Kumar Srivastava, Senthil Kumar Mohan “Defensive Modelling of Fake News Through Online Social Networks”, — Online social networks (OSNs) IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS
- DeBeer, Dylan &Matthee,Machdel,” Approaches to Identify FakeNews: A Systematic Literature Review.”: Integrated Science in Digital Age2020. [3], Bandar Alghamdi, Fahad Alharbi,” An Intelligent Model for Online Recruitment Fraud Detection”. Journal of Information Security. (2019)
- Pham, Trung Tin” A Study on Deep Learning for Fake News Detection.” Journal of Information Security. (2019)
- Manoj Kumar Balwant.” Bidirectional LSTM Based on POS tags and CNN Architecture for Fake News” 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). (2019)
- Amjad, Maaza,Sidorov, Grigoria, Zhila, Alisaa, Gómez-Adorno, Helenab, Voronkov, Iliac, Gelbukh, Alexander.” Bend the truth” Special section: Selected papers of LKE 2019 In [7], Rami Mohawesh, Son Tran, Robert Ollington, Shuxiang Xu” Analysis of concept drift in fake reviews detection.” Expert Systems with Applications. (2021)