International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Accurate prediction of real or fake job postings using Machine learning

Dr P M JOE PRATAP
Mentor, Professor, Department of Information Technology,
R.M.D Engineering College, Tamil Nadu, India

SHAJITH BOBBY K R, M.SRIJAYABALAJI, C.N.A.YOGESH
Author, Student, Department of Information Technology,
R.M.D Engineering College, Tamil Nadu, India

Abstract

This document with everything is being online now it allows people to reduce their manual efforts since all of the job postings are posted online now so it gives the company a wide range of area to gather the talented candidates and for the people who are searching they can also know the information most companies can directly post the job. All job postings which are posted are not true there are fraudulent job postings. So we try to classify the fraudulent posting from real. The aim is to predict machine learning based techniques for real or fake job prediction results in best accuracy. The analysis of dataset is done by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset.

Keywords: Dataset, Machine learning-Classification methodology, python, Prediction of Accuracy result.

1. Introduction

1.1 Domain Overview: Machine learning could be a form of AI (AI) that has computers with the power to find out while not being expressly programmed. Machine learning focuses on the event of Programs that may have modification once exposed to new knowledge. Machine learning(ML) is that the study of algorithms that improve through by the utilization of information.[1] it is a section of AI. Machine learning algorithms build a model supported sample knowledge, referred to as "training data", so as to create predictions while not being expressly programmed .[2] Machine learning algorithms square measure utilized in a good style of applications, like email filtering and vision. Data processing a field of study, that specialize in exploratory knowledge analysis through unattended learning.

1.2 Objectives: The goal is to develop a machine learning model for accurate prediction of real and fake job, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

1.3 Drawback Statement: [3] There are plenty of job advertisements on the web, even on the purported job advertising sites, that ne'er appear faux. however when the choice, the supposed recruiters begin soliciting for the cash and therefore the bank details. several of the candidates fall in their entice and lose plenty of cash and therefore the current job generally. So, it's higher to spot whether or not employment advert announce on the location is real or faux. distinguishing it manually is incredibly tough and nearly not possible.

1.4 Scope: the problems of faux job postings ar typically related to Technical platforms and several other such platforms are attempting their best to spot scams. but it's much not possible to try and do it manually, thence AI is employed to resolve the matter. Though, at the start making such a system needs great deal of resources, it'll resolve the burden of assessing the genuineness of each job posting out there on the platform.

2.1 Literature Survey:

Title:1 A Sensitive Stylistic Approach to Identify Fake News on Social Networking

Author: Nicollas R. de Oliveira, Dianne S. V. Medeiros

This paper presented a computational analysis, based on natural language processing, efficiently applying unsupervised learning algorithms, such as one- class SVM, in detecting fake news in texts extracted from social media. They proposed to apply to original data both dimensionality reduction technique, through latent semantic analysis (LSA), and data compaction through our proposed methodologies. Three different news classification methodologies were implemented – two

employing cascading or unique configurations of learning algorithms and the other statistically evaluating the difference between the types of news. They proposed analysis based on natural language processing, efficiently applying machine learning algorithms to detect fake news in texts extracted from social media. The analysis considers news from Twitter, from which approximately 33,000 tweets were collected, assorted between real and proven false. In assessing the quality of detection, 86% accuracy, and 94% precision stand out even employing a dimensional reduction to one-sixth of the number of original features.

Title:2 pretend Job achievement Detection exploitation Machine Learning Approach

Author: ShawniDutta , Samir Kumar Bandyopadhyay

Employment scam detection can guide job-seekers to induce solely legitimate offers from corporations. For effort employment scam detection, many machine learning algorithms area unit projected as countermeasures during this paper. supervised mechanism is employed to exemplify the utilization of many classifiers for employment scam detection. To avoid dishonest post for job within the web, an automatic tool exploitation machine learning based mostly classification techniques is projected within the paper. totally different classifiers area unit used for checking dishonest post within the internet and therefore the results of these classifiers area unit compared for characteristic the most effective employment scam detection model. It indicates that ensemble classifiers area unit the most effective classification to notice scams over the only classifiers.

Title:3 Spammer Detection and Fake User Identification on Social Networks

Author: FAIZA MASOOD , GHANA AMMAD, AHMAD ALMOGREN, ASSAD ABBAS , HASAN ALI KHATTAK

They performed a review of techniques used for detecting spammers on Twitter. In addition, we also presented taxonomy of Twitter spam detection approaches and categorized them as fake content detection, URL based spam detection, spam detection in trending topics, and fake user detection techniques. We also compared the presented techniques based on several features, such as user features, content features, graph features, structure features, and time features. Moreover, the techniques were also compared in terms of their specified goals and datasets used. It is anticipated that the presented review will help researchers find the information on state-of-the-art Twitter spam detection techniques in a consolidated form. Despite the development of efficient and effective approaches for the spam detection and fake user identification on Twitter, there are still certain open areas that require considerable attention by the researchers. The issues are briefly highlighted as under: False news identification on social media networks is an issue that needs to be explored because of the serious repercussions of such news at individual as well as collective level .Another associated topic that is worth investigating is the identification of rumor sources on social media.

3.1 Existing System:

A mathematical model to check the dynamic unfolding and dominant activities of message transmission in OSN was projected The projected model employs differential equations for work the impact of verification and obstruction of users and therefore the spread of messages on OSNs. It describes however info gets disseminated among teams with the influence of various info refuting measures. This model is predicated on differing types of epidemic categories and has 2 layers of management mechanism to manage the rumor within the social network.[4] This model assumes that every one users ar prone which means anyone could flip a victim of info or untrusted message. for defense, initially, the users ar attested employing a employing a. Hence,before acceptive the request of any user, the user authentication technique is applied, and therefore the dependability of the messages from this user is evaluated so as to reduce the activities of malicious users to the OSN If the worth of R0 is a smaller amount than one (R0 < 1), then pretend message spreading within the on-line network won't be distinguished, otherwise if R0 > one the rumor can act the OSN.

Drawbacks:

This is just a mathematical model it does not classify the fake news.

Accuracy, Recall F1 score metrics are not calculated and machine learning algorithms are not applied.

3.2. Proposed System:

The proposed method is built a machine learning model to classify the real or fake job posting to overcome this method to implement machine learning approach by user interface of GUI application. The dataset is first preprocessed and the columns are analyzed to see the dependent and independent variable and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

Feasibility Study:

1. Information Wrangling:

In this section of the report can load within the information ,check for cleanliness, and so trim and clean given dataset for analysis. certify that the document steps fastidiously and justify for cleansing selections.

2. Information collection:

The information set collected for predicting given data is split into coaching set and take a look at set. Generally, 7:3 ratios area unit applied to separate the coaching set and take a look at set. the info Model that was created mistreatment Random Forest, logistic, call tree algorithms, K-Nearest Neighbor (KNN) and Support vector classifier (SVC) area unit applied on the coaching set and supported the take a look at result accuracy, take a look at set prediction is finished.

3. Preprocessing:

The data that was collected may contain missing values that will cause inconsistency. to realize higher results information ought to be preprocessed therefore on improve the potency of the formula. The outliers need to be removed and conjointly variable conversion ought to be done.

4. Building the classification model:

The predicting the air quality downside, call tree formula prediction model is effective due to the subsequent reasons: It provides higher ends up in classification downside.

It is sturdy in preprocessing outliers, immaterial variables, and a combination of continuous, categorical and distinct variables.

It produces out of bag estimate error that has evidenced to be unbiased in several tests and it's comparatively simple to tune with.

5. Construction of a prophetical Model:

Machine learning desires information gathering have heap of past data's. information gathering have spare historical information and data. Before information pre-processing, data can't be used directly. It's accustomed preprocess then, what reasonably formula with model. coaching and testing this model operating and predicting properly with minimum errors. Tuned model concerned by tuned time to time with up the accuracy.

Exploratory knowledge Analysis:

Multiple datasets from completely different sources would be combined to make a generalized dataset, so completely different machine learning algorithms would be applied to extract patterns and to get results with most accuracy.

Advantages:

These reports ar to the investigation of pertinence of machine learning techniques for job posting classification.

It highlights some observations on future analysis problems, challenges, and needs.

System Implementation

Modules:

1. Data validation and pre-processing technique.
2. Data visualization and training a model by given attributes.
3. Performance measurements of ML algorithms.
4. Implementation of LSTM model for coaching and testing.

1 Data validation and pre-processing technique.

Importing the library packages with loading given dataset. To analyzing the variable identification by information form, information kind and evaluating the missing values, duplicate values. information improvement / making ready by rename the given dataset and drop the column etc. to research the uni-variate, bi-variate and multi-variate method. The steps and techniques for information improvement can vary from dataset to dataset. the first goal of information improvement is to discover and take away errors and anomalies to extend the worth of information in analytics and deciding.

2 Data visualization and training a model by given attributes.

Data mental image is a very important talent in applied statistics and machine learning. Statistics will so concentrate on quantitative descriptions and estimations of knowledge. knowledge mental image provides a very important suite of tools for gaining a qualitative understanding. this could be useful once exploring and attending to understand a dataset and may facilitate with characteristic patterns, corrupt knowledge, outliers, and far a lot of. With a bit domain data, knowledge visualizations is accustomed specific and demonstrate key relationships in plots and charts

3 Performance measurements of ML algorithms:

It is a statistical method for Associate in Nursing analyzing a information set throughout that there unit one or further freelance variables that verify associate outcome. the top result's measured with a divided variable (in that there unit alone a pair of possible outcomes). The goal of supplying regression is to look out the foremost effective fitting model to clarify the link between the divided characteristic of interest (dependent variable = response or outcome variable) and a set of freelance (predictor or explanatory) variables. supplying regression may well be a Machine Learning classification formula that is accustomed predict the possibility of a categorical variable.

True Positive Rate = TP /(TP+FN)

False Positive rate = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non- defaulters.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are literally correct.

exactness = TP / (TP + FP)

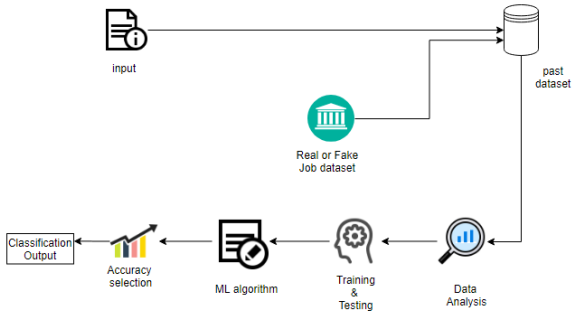
Precision is that the magnitude relation of properly foretold positive observations to the full foretold positive observations. The question that this metric answer is of all passengers that tagged as survived, what percentage really survived? High exactness relates to the low false positive rate. we've got zero.788 exactness that is pretty smart.

Recall: The proportion of positive determined values properly foretold. (The proportion of actual defaulters that the model can properly predict)
Recall = TP / (TP + FN)

4 Implementation of LSTM model for coaching and testing:

Long Short Term Memory networks – sometimes simply referred to as "LSTMs" – square measure a special reasonably RNN, capable of learning semipermanent dependencies. They were introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many of us. They work enormously well on an oversized kind of issues, and square measure currently wide used. LSTMs square measure expressly designed to avoid the semipermanent dependency drawback. memory data for long periods of your time is much their default behavior, not one thing they struggle to learn!. LSTMs even have this chain like structure, however the continuation module incorporates a completely different structure. rather than having one neural network layer, there square measure four, interacting in a {very} very special method. The key to LSTMs is that the cell state, the horizontal line running through the highest of the diagram. The cell state is reasonably sort of a conveyer belt. It runs straight down the complete chain, with just some minor linear interactions. It's terribly straightforward for data to simply flow on it unchanged. The LSTM will have the power to get rid of or add data to the cell state, rigorously regulated by structures referred to as gates.Gates square measure the way to optionally let data through. they're composed out of a sigmoid neural internet layer and a pointwise multiplication operation. The sigmoid layer outputs numbers between zero and one, describing what quantity of every element ought to be let through. a price of zero means that "let nothing through," whereas a price of 1 means that "let everything through!"

System Architecture:



Future work:

To automate this process by show the prediction result in web application or desktop application.

To optimize the implementation of Artificial Intelligence environment.

Conclusion:

There are a lot of job advertisements on the internet, even on the reputed job advertising sites, which never seem fake. But after the selection, the so-called recruiters start asking for the money and the bank details. Many of the candidates fall in their trap and lose a lot of money and the current job sometimes. So, it is better to identify whether a job advertisement posted on the site is real or fake. Identifying it manually is very difficult and almost impossible. The Issues of fake job postings are often associated with Technical platforms and several such platforms are trying their best to identify scams. However it is practically impossible to do it manually, hence AI is used to solve the problem. Though, initially creating such a system requires large amount of resources, It will resolve the burden of assessing the authenticity of every job posting available on the platform.

References:

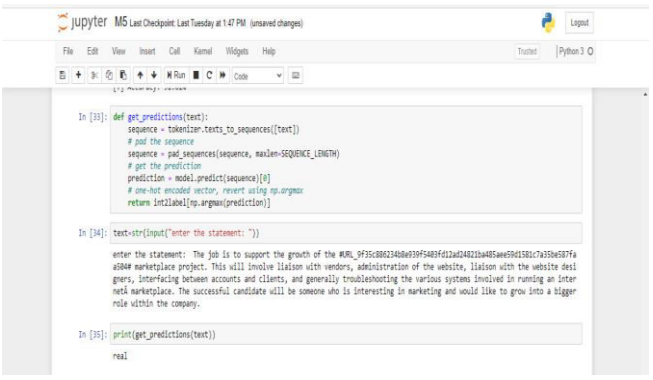
[1] S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, and W. Jia, "Modelingpropagation dynamics of social network worms," IEEE Trans. Parallel Distrib.
Syst., vol. 24, no. 8, pp. 1633–1643, Aug. 2013.

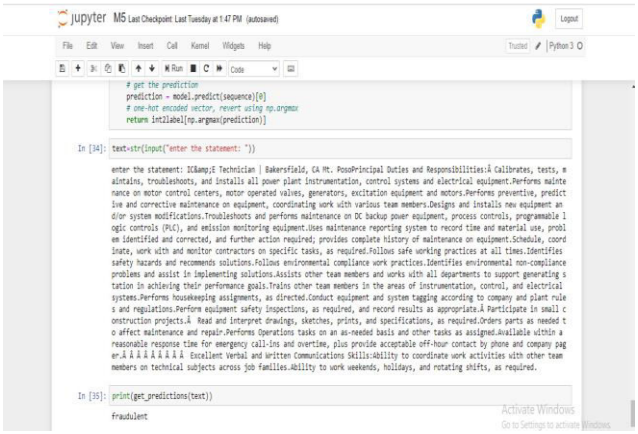
[2] E. Lebensztayn, F. P. Machado, and P. M. Rodríguez, "On the behaviour of a rumour technique with random stifling," Environ. Model.Softw., vol. 26, no. 4, pp. 517–522, Apr. 2011.

[3] L. Li et al., "Characterizing the propagation of situational knowledge in social media throughout COVID-19 epidemic: A case study on weibo," IEEE Trans. Comput. Social Syst., vol. 7, no. 2, pp. 556–562, Apr. 2020.

[4] S. Sommariva, C. Vamos, A. Mantzarlis, L. U.-L. Dào, and D. Martinez Tyson, "Spreading the (fake) news: exploring health messages on social media and additionally the implications for health professionals using a case study," Amer. J. Health Edu., vol. 49, no. 4, pp. 246–255, Jul. 2018.

Screenshots Real or Fake prediction




jupyter M5 Last Checkpoint: Last Tuesday at 1:47 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted Python 3
Run Code
# get the prediction
prediction = model.predict(sequence)[0]
# one-hot encoded vector, revert using np.argmax
return int2label[np.argmax(prediction)]
In [34]: text=str(input("enter the statement: "))
enter the statement: IC&E Technician | Bakersfield, CA Rt. PosoPrincipal Duties and Responsibilities:Â Calibrates, tests, maintains, troubleshoots, and installs all power plant instrumentation, control systems and electrical equipment.Performs maintenance on motor control centers, motor operated valves, generators, excitation equipment and motors.Performs preventive, predictive and corrective maintenance on equipment, coordinating work with various team members.Designs and installs new equipment and/or system modifications.Troubleshoots and performs maintenance on DC backup power equipment, process controls, programmable logic controls (PLC), and emission monitoring equipment.Uses maintenance reporting system to record time and material use, problem identified and corrected, and further action required; provides complete history of maintenance on equipment.Schedule, coordinate, work with and monitor contractors on specific tasks, as required.Follows safe working practices at all times.Identifies safety hazards and recommends solutions.Follows environmental compliance work practices.Identifies environmental non-compliance problems and assist in implementing solutions.Assists other team members and works with all departments to support generating station in achieving their performance goals.Trains other team members in the areas of instrumentation, control, and electrical systems.Performs housekeeping assignments, as directed.Conduct equipment and system tagging according to company and plant rules and regulations.Perform equipment safety inspections, as required, and record results as appropriate.Â Participate in small construction projects.Â Read and interpret drawings, sketches, prints, and specifications, as required.Orders parts as needed to affect maintenance and repair.Performs Operations tasks on an as-needed basis and other tasks as assigned.Available within a reasonable response time for emergency call-ins and overtime, plus provide acceptable off-hour contact by phone and company pager.Â Â Â Â Â Â Â Â Â Â Â Â Excellent Verbal and Written Communications Skills:Ability to coordinate work activities with other team members on technical subjects across job families.Ability to work weekends, holidays, and rotating shifts, as required.
In [35]: print(get_predictions(text))
fraudulent
Activate Windows