CONSULT JOB

Real/fake job posting prediction

Godala Soma Sandeep Reddy
(20103059)
Gourav Gujariya
(20103060)
Lovepreet Kumar
(20103084)

03.04.2022

Machine learning

ABSTRACT-Unfortunately, fraudulent job postings are still prevalent online as scammers try to take advantage of impatient job seekers.Job scams are becoming more and more common and job seekers continue to be the target of insidious scammers. According to theBetter Business Bureau, about 14 million people are involved in workplace fraud each year, and people who are not familiar with workplace fraud are more likely to lose money because of it. If you are looking for a new job, beware of job scams by learning what to watch out for in order to better protect yourself from them.
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 text 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.

KEYWORDS-Decision tree classification,Pipeline,logisticRegression, Random forest naive bayes, support vector machine,job and career and other.

INTRODUCTION AND MOTIVATION-

Nowadays to reach a mass number of people, many companies are preferring to advertise their hiring process online. Though it has a huge advantage, scammers misuse it by using a reputed company's name and take money from the job applicants. These fraudulent job advertisements draws good attention from the applicants. By these fraudulent job advertisements, many applicants are losing their money and also the companies are losing their reputation.Because of these fake adds ,people are in confusion whether an advertisement is fake or real.Many people are losing opportunities due to this confusion.

To avoid these we can use a machine learning approach to identify the fraud job advertisements. We can classify whether a job advertisement is fake or real so that applicants can identify real advertisements and can apply to the jobs.It saves a lot of time to the applicants.This classification also helps to control the spammers.Many companies are introducing this detection application to help the applicants.A.As we are classifying that if a job is fraud or not so The following classifiers are used while detecting fake job posts

a. Naive Bayes Classifier:The Naive Bayes classifier is a supervised classification tool that exploits the concept of Bayes Theorem of Conditional Probability. The decision made by this classifier is quite effective because I work as if the point belongs to this field if the feature is already in favor or we can say that it uses Bayes optimal classifier. This classifier obtains a very promising result in the following scenario- when the features are independent or features are completely functionally dependent. The accuracy of this classifier is not related to feature dependencies;
rather it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy .

b) Multi-Layer Perceptron Classifier:Multi-layer perceptron can be used as a supervised classification tool by incorporating optimized training parameters. For a given problem, the number of hidden layers in a multilayer perceptron and the number of nodes in each layer can differ. The decision of choosing the parameters depends on the training data and the network architecture .

c) K-nearest Neighbor Classifier:K-Nearest Neighbour Classifier , often known as lazy learners, identifies objects based on closest proximity of training examples in the feature space. The classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on choosing the appropriate value of k .

d) Decision Tree Classifier:A Decision Tree (DT) is a classifier that exemplifies the use of tree-like structure. It gains knowledge on classification. Each target class is denoted as a leaf node of DT and non-leaf nodes of tree.

e)support vector machine:support vector machine is also a supervised machine learning algorithm and is used for classification problems it uses kernel or we can say plain for some of the problem and work with the distance between points and create a plain,line or hyper plain for there classification. It also work fine with big dataset and use vector for classification.

f)Random forest:random is similar to decision tree which work with the tree but it create random trees and do the computation of every possible computation for good prediction and in random forest the chances of creating a good classification model are more with high accuracy cause it create a large no of trees and branches with many computations.

g)Pipeline:pipeline is for the transformation the file work into ease for next time cause when ever new data to be entered or used in the model it will go through a following pipeline or we can say that through a follofwing process fo data processing and cleaning so that when it reach model it is ready to use.

h)we have also used some natural language processing modules to work with the headlines of the fake jobs and real jobs and convert them into int to make model work easy.

PREVIOUS WORKS

A. Review Spam Detection- People often post their reviews online forum regarding the regarding their previous job or interview experience. It may guide other new job seekers while applying for a job. In this context, spammers can manipulate reviews for gaining profit and hence it is required to develop techniques that detects these spam reviews. This can be implemented by extracting features from the reviews by extracting features using Natural Language Processing (NLP).

- **MACHINE LEARNING APPROACH:** We split review centric features into several categories. First, we have bag-of-words(In a bag of words approach, individual or small groups of words from the text are used as features. These features are called n-grams and are made by selecting n contiguous words from a given sequence, i.e., selecting one, two or three contiguous words from a text. These are denoted as a unigram, bigram, and trigram (n = 1, 2 and 3) respectively.) , and bag-of-words combined with term frequency features. Next, we have Linguistic Inquiry and Word Count (LIWC) output, parts of speech (POS) tag frequencies, Stylometric and Syntactic features. Finally, we have review characteristic features that refer to information about the review not extracted from the text.

B. Email Spam Detection- Unwanted bulk mails, belong to the category of spam emails, often arrive to user mailbox. This may lead to unavoidable storage crisis as well as bandwidth consumption. To solve this problem, various big companies like Gmail, Yahoo mail and Outlook service providers incorporate spam filters using Neural Networks. While addressing the problem of email spam detection, content based filtering, case based filtering, heuristic based filtering, memory or instance based filtering, adaptive spam filtering approaches are taken into consideration.

MACHINE LEARNING APPROACH: on the basis of following points we built our model of email spam detection :

- **Content Based Filtering Technique:** This method normally analyses words, the occurrence, and distributions of words and phrases in the content of emails and used then use generated rules to filter the incoming email spams.
- **Heuristic or Rule Based Spam Filtering Technique:** This approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message. Several similar patterns increase the score of a message. In contrast, it deducts from the score if any of the patterns did not correspond. Any message's score that surpasses a specific threshold is filtered as spam; else it is counted as valid.
- **Previous Likeness Based Spam Filtering Technique:** This approach uses memory-based, or instance-based, machine learning methods to classify incoming emails based to their resemblance to stored examples (e.g. training emails).
- **Adaptive Spam Filtering Technique:** The method detects and filters spam by grouping them into different classes. It divides an email corpus into various groups, each group has an emblematic text. A comparison is made between each incoming email and each group, and a percentage of similarity is produced to decide the probable group the email belongs to.

C. Fake News Detection- Fake news in social media as spread through fake accounts using certain keywords repeatedly to attract target audience. The fundamental study of fake news detection relies on three perspectives- how fake news is written, how fake news spreads, how a user is related to fake news. Features related to news content and social context are extracted and a machine learning models are imposed to recognize fake news.

MACHINE LEARNING APPROACH: using following techniques we built our model of email spam detection :

- **TfidfVectorizer :**

TF (Term Frequency): The number of times a word appears in a document is its Term Frequency. A higher value means a term appears more often than others, and so, the document is a good match when the term is part of the search terms.

- **IDF (Inverse Document Frequency):** Words that occur many times a document, but also occur many times in many others, may be irrelevant. IDF is a measure of how significant a term is in the entire corpus.
- The TfidfVectorizer converts a collection of raw documents into a matrix of TF-IDF features.

- **PassiveAggressiveClassifier :** Passive Aggressive algorithms are online learning algorithms. Such an algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting. Unlike most other algorithms, it does not converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the weight vector.

REFERENCES

- Real / Fake Job Posting Prediction on #kaggle via @KaggleDatasets https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction?utm_medium=social&utm_campaign=kaggle-dataset-share&utm
- https://towardsdatascience.com/fake-job-predictor-a168a315d866
- https://www.researchgate.net/profile/Samir-Bandyopadhyay/publication/341325717_Fake_Job_Recruitment_Detection_Using_Machine_Learning_Approach/links/5ebad816299bf1c09ab91341/Fake-Job-Recruitment-Detection-Using-Machine-Learning-Approach.pdf/