Detecting Fake Job Posting Using NLP

BHOOMIKA AGRAHARI

DATE: 18/12/2023

1. Abstract:

Within the job market, plenty of juniors fall prey to fake job postings which can lead them to be scammed and lose interest in the field, or in the worst case, lose hope in their dreams or even get caught up in criminal acts. The problem regarding this is prevalent within the domain. With how ambiguous the data science professional world has been with its job titles, it has also led to a plethora of people, junior and experienced alike, ending up misunderstanding a job posting and getting a role they didn't have in mind. This project aims to minimize such occurrences.

2. Problem Statement:

As individuals navigate the job market, they often face the potential pitfalls of fraudulent job advertisements, which may lead to exploitation or scamming by deceptive entities. There's also the risk of aligning with companies that lack a comprehensive understanding of the roles they're advertising.

This project aims to allow people to glean from the job posting whether it is a fake job or also a true one that shouldn't be trusted. It also has as a goal to identify the patterns of words perceived within these aforementioned job postings that should be avoided.

3. Market/Customer/Business Need Assessment

The job market needs a reliable tool to identify fake job postings, safeguarding job seekers from scams and misinformation. The tool should provide clarity on job descriptions and help individuals make informed decisions about job applications. By conducting a these comprehensive, the fake job predicting software can be tailored to address the critical concerns and requirements of both job seekers and recruitment platforms, ensuring its relevance and impact in the industry.

1. *Job Seeker Concerns:*

- Identify the growing concern among job seekers regarding fake job postings.
- Recognize the impact of falling victim to fraudulent job advertisements on job seekers' trust and confidence.

2. *Industry Reputation:*

- Assess the impact of fake job postings on the reputation of the recruiting and job search industry.
- Recognize the need for solutions that can enhance the credibility of job platforms.

3. *Financial Losses:*

- Understand the financial losses incurred by individuals who fall prey to fake job scams.
- Recognize the need for a solution that protects job seekers from financial exploitation.

4. *Recruitment Platforms' Responsibility:*

- Evaluate the responsibility of recruitment platforms to provide a secure and trustworthy job-seeking environment.
- Recognize the potential legal and reputational consequences for platforms associated with fake job postings.

5. *User Experience:*

- Evaluate the impact of fake job postings on the overall user experience of job seekers.
- Recognize the importance of a positive and secure user experience for the sustained growth of job platforms.

4. Target Specifications and Characterization

1. *Accuracy:*

- Achieve a high level of accuracy in identifying fake job postings.
- The product should aim for an accuracy rate that minimizes false positives and false negatives, ensuring reliable detection.

2. *Scalability:*

- Provide scalable solutions to accommodate a large volume of job postings.
- The product should be capable of handling the increasing number of job listings on diverse platforms.

3. *Real-Time detection:*

- Enable real-time detection of fake job postings as soon as they are posted.
- The product should have minimal latency in analyzing and flagging potential fake job advertisements.

4. *Adaptability:*

- Adapt to evolving techniques used by scammers to create fake job postings.
- Implement machine learning models that continuously learn and adapt to new patterns of fraudulent job listings.

5. *User-Friendly Interface:*

- Provide an intuitive and user-friendly interface for both job seekers and platform administrators.
- Design a dashboard or plugin that is easy to navigate and understand, ensuring accessibility for users with varying technical expertise.

6. *Integration Capabilities:*

- Integrate seamlessly with popular job platforms and recruitment websites.
- Develop APIs or plugins that can be easily integrated into existing recruitment platforms, making adoption straightforward.

5. External Search (information sources/references)

The dataset I used is from kaggle, you can easy access it using this link (https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction/data)

This dataset contains 18K job descriptions out of which about 800 are fake. The data consists of both textual information and meta-information about the jobs. The dataset can be used to create classification models which can learn the job descriptions which are fraudulent.

The dataset is very valuable as it can be used to answer the following questions:

1. Create a classification model that uses text data features and meta-features and predict which job description are fraudulent or real.
2. Identify key traits/features (words, entities, phrases) of job descriptions which are fraudulent in nature.
3. Run a contextual embedding model to identify the most similar job descriptions.
4. Perform Exploratory Data Analysis on the dataset to identify interesting insights from this dataset.

   Let's view the dataset:

In [3]:
```
#Explore first five rows
train_df.head()
```

Out[3]:

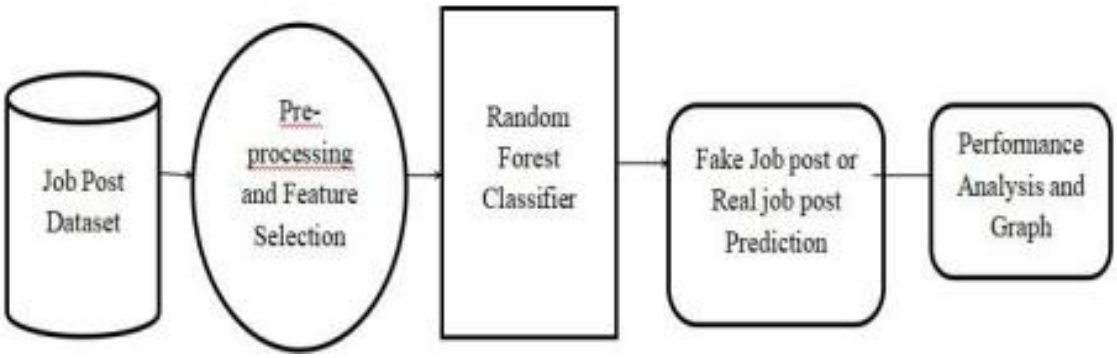
	job_id	title	location	department	salary_range	company_profile	description	requirements
0	1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we expect from you:Your key responsibilit...
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...
3	4	Account Executive - Washington DC	US, DC, Washington	Sales	NaN	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...
4	5	Bill Review Manager	US, FL, Fort Worth	NaN	NaN	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...

6. Applicable Regulations

1. Must provide access to the 3rd party websites to audit and monitor the authenticity and behavior of the service.
2. Enabling open-source, academic and research community to audit the Algorithms and research on the efficacy of the product.
3. Laws controlling data collection: Some websites might have a policy against collecting customer data in form of reviews and ratings.
4. Must be responsible with the scraped data: It is quintessential to protect the privacy and intention with which the data was extracted.

7. Business Model

The project is to find the phony jobs to avoid users getting into the scams. This makes assurance that the data they provide at the time of recruitment will not be misused. We are working on a EMSCAD dataset to find better results using different algorithms. The dataset for fake job post is collected and preprocessed. The feature selection is the process of selecting some important features from the data required for analyzing and getting a proper output. We are applying the Random Forest Classifier to detect whether the job posted is a fake or a legitimate one.



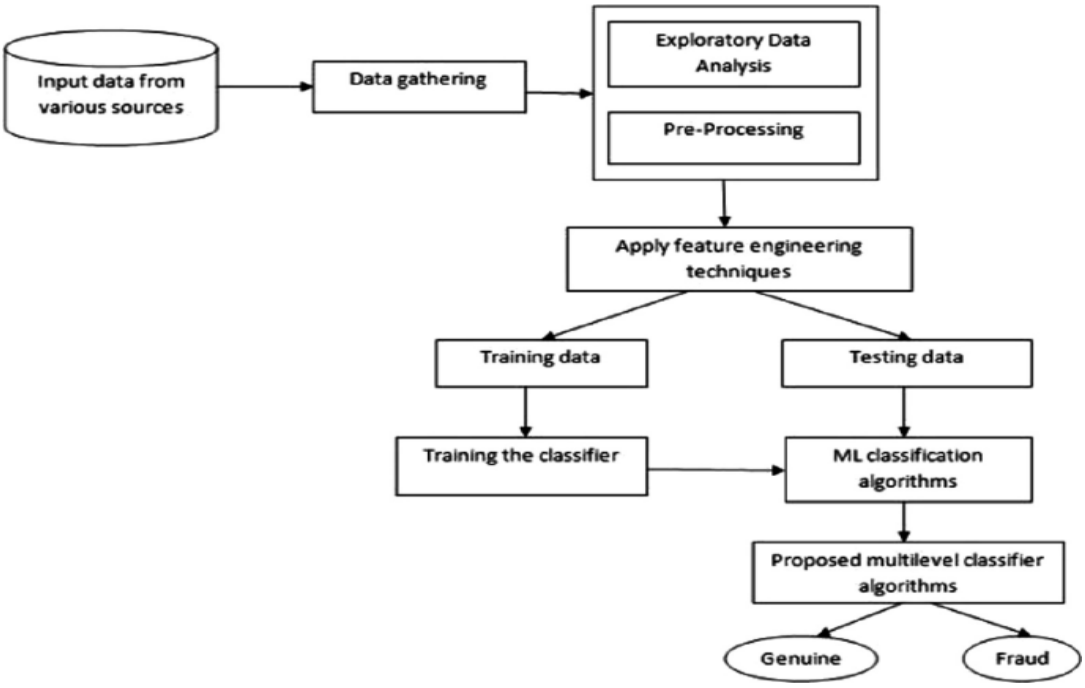
8. Concept Generation

This product requires the tool of machine learning models to be written from scratch in order to suit our needs. Tweaking these models for our use is less daunting than coding it up from scratch. A well trained model can either be repurposed or built. But building a model with the resources and data we have is dilatory but possible. The customer might want to spend the least amount of time giving input data. This accuracy will take a little effort to nail, because it's imprudent to rely purely on Classic Machine Learning algorithm.

```
In [56]: from sklearn.ensemble import RandomForestClassifier
         rfc = RandomForestClassifier(n_jobs=3,oob_score=True,n_estimators=100,criterion="entropy")
         model = rfc.fit(X_train,y_train)

In [58]: pred = rfc.predict(X_test)
         score = accuracy_score(y_test,pred)
         score

Out[58]: 0.9735272184936614

In [60]: print('Classification Report\n')
         print(classification_report(y_test,pred))
         print("Confusion Matrix\n")
         print(confusion_matrix(y_test,pred))

         Classification Report

                       precision    recall  f1-score   support

                    0       0.97      1.00      0.99      5117
```

9. Final Product Prototype (abstract) with Schematic Diagram

- The product will be GUI based web Application in which user just have to enter the job details in a paragraph like manner and our model will be able to extract useful information from that set of data and give its prediction according to that.



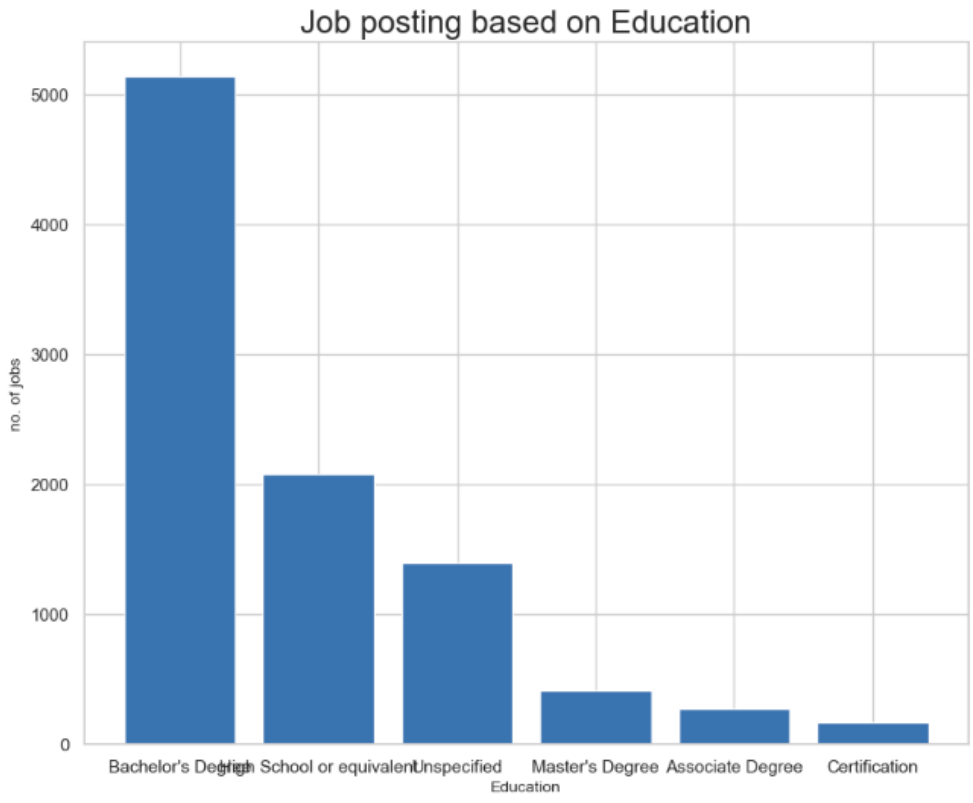
- In this project we have been do a lot of things from data gathering to Exploratory data analysis and pre-processing then applying feature engineering techniques etc.
- Next we have done some visualizations to understand our data more efficiently using various graphs Such as word cloud, bar graph etc.

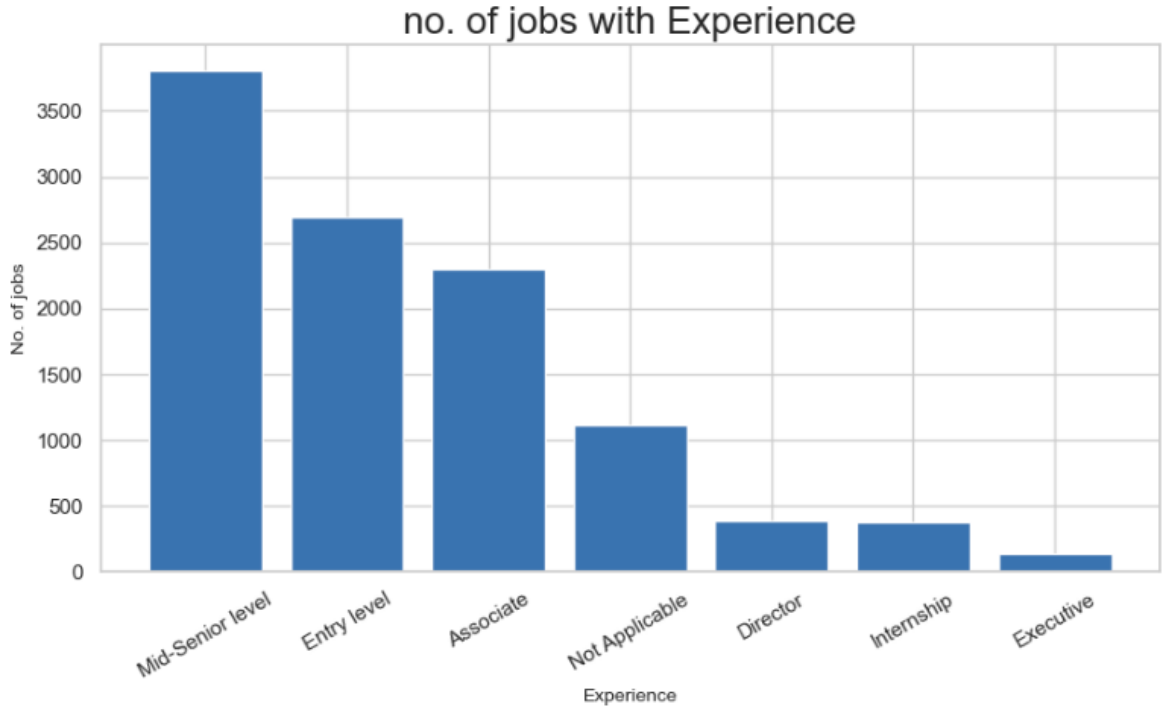
10. Code Implementation/Validation on Small Scale

The code Is implemented using python, pandas and some machine learning algorithm using which I have trained the data. The data has been gathered from kaggle which is the easy source of data being used these days. I have use all the machine learning pipeline which is used to develop a machine learning application.

Some visualization of the data

```
In [39]: plt.figure(figsize=(10,8))
         plt.title('Job posting based on Education',size=20)
         plt.bar(edu.keys(),edu.values())
         plt.ylabel('no. of jobs',size=10)
         plt.xlabel('Education',size=10)
Out[39]: Text(0.5, 0, 'Education')
```





Some EDA Operations performed on data

In [43]:
```
df['text']=df['title']+' '+df['company_profile']+' '+df['description']+' '+df['requirements']+' '+df['benefits']
del df['title']
del df['location']
del df['department']
del df['company_profile']
del df['description']
del df['requirements']
del df['benefits']
del df['required_experience']
del df['required_education']
del df['industry']
del df['function']
del df['country']
```

In [44]:
```
df.head()
```

Out[44]:

	fraudulent	text
0	0	Marketing Intern We're Food52, and we've creat...
1	0	Customer Service - Cloud Video Production 90 S...
2	0	Commissioning Machinery Assistant (CMA) Valor ...
3	0	Account Executive - Washington DC Our passion ...
4	0	Bill Review Manager SpotSource Solutions LLC i...

In [49]:
```
punctuation = string.punctuation

nlp = spacy.load("en_core_web_sm")

stop_words = spacy.lang.en.stop_words.STOP_WORDS

parser = English()

def spacy_tokenizer(sentence):
    mytokens = parser(sentence)

    mytokens = [word.lemma_.lower().strip() if word.lemma_ != "-PRON-" else word.lower_ for word in mytokens]
    mytokens = [word for word in mytokens if word not in stop_words and word not in punctuations]

    return mytokens

class predictors(TransformerMixin):
    def transform(self,X, **transform_params):
        return [clean_text(text) for text in X]

    def fit(self,X,y=None, **fit_params):
        return self
    def get_params(self,deep=True):
        return {}

def clean_text(text):
    return text.strip().lower()
```

Github Link: https://github.com/bhoomika297/feynn-labs/upload/main

11. Conclusion

The detection of job scams has recently become a major problem worldwide. We have examined the effects of employment scams in this project since they might be a very lucrative topic of study and make it difficult to identify the posts of fake job. We made use of EMSCAD dataset, which includes real-time job postings. Random forest classifier gives 97 percent of accuracy then the previously used algorithms like SVM, Decision tree classifier, etc which gives 90 percent of accuracy. We are making the hiring procedure through online safer, by avoiding frauds and scams in the job.

Therefore, you can go for applying the jobs through online process. Therefore, avoiding the financial losses of a person and protecting the personal information of a person.