Detecting Fake Companies on LinkedIn: A Machine Learning Approach

Kaushal Thaker

kaushalthaker145@gmail.com

Abstract. The proliferation of fake companies on professional networking platforms like LinkedIn poses significant risks to users, including scams, identity theft, and data breaches. To address this issue, I propose a novel approach utilizing machine learning techniques to detect the legitimacy of companies on LinkedIn. I conducted extensive research to identify key factors influencing company legitimacy and incorporated them into a comprehensive dataset. Employing popular machine learning algorithms such as Support Vector Machine, Decision Tree, and K-Nearest Neighbor, I trained models to predict whether a company is real or fake. Additionally, I developed a risk labeling system to assess the level of risk associated with each company. My approach includes the integration of a scrapper component to extract essential information from resumes while ensuring user privacy and security. Through rigorous analysis and evaluation, I demonstrate the effectiveness of My methodology in accurately identifying fake companies and reducing risks for LinkedIn users. This research contributes valuable insights and practical solutions to enhance trust and security in professional networking environments.

1 Introduction

Detecting fake companies on LinkedIn is important because it helps to protect users from scams and identity theft. Scammers can easily access and gather personal and sensitive user information, and users are often less aware and least concerned about security settings. Fake profiles can have an adverse effect on the trustworthiness of the network as a whole and can represent significant costs in time and effort in building a connection based on fake information.

I recently encountered a situation where a company posted a job opening for software developers. During the final stage of the application process, the recruitment team asked the candidates to write a review of the products they sold and give ratings. However, later, the candidate learned that it was a fake company. Job openings for software developers typically do not involve reviewing and providing ratings for a company's products. This might not look like a major issue, but be aware that a few companies tend to go further in scam and ask the candidate to pay for the courses they offer as a part of the recruitment process.

Most of the time, I provide My resumes to fake companies, which steal My information and use it to commit fraud. There is even a data breach instance

where LinkedIn has already faced data breaches. One of the fake profile companies obtained information about users from resumes sent to them.

In 2019, a fake company called "Asheville Matrix LLC" posted job listings for various positions on LinkedIn. The scammers conducted interviews and even extended offers to unsuspecting candidates. However, the company did not exist, and the scammers attempted to collect personal information and potentially engage in identity theft.

I had seen an instance in one of the user's LinkedIn posts where he had a screenshot of the mail that he received from xyz company. The email states that he was selected for a job interview. The message says that they obtained or her resume from Indeed.com. But the actual fact is, that particular candidate never applied to this job from Indeed.com. The sentence formation in the data provided in the mail was also not that good. The email reference of the recruiter person mentioned in the mail had a generic expansion for the email that is gmail.com and not company specific email. The phone number provided was also VOIP (voice over internet protocol) based, and is "non-fixed". Using VOIP, you can get a new phone number any time you choose; difficult to investigate too.

This provided us motivation to look into companies that are not legitimate, and I wanted to build a model where it detects whether the company is real or fake. On the other hand, the most important privacy concern here is that no individual wants to give away his resume without knowing the legitimacy of the company because if in case the resume goes into the hands of fake companies, then they can do any scams like sending phishing emails to a candidate or be a cause of user data breach.

My major goal was not to give away My resume to the company without knowing how legitimate the company was. Each company might have different levels of risks associated with it. So, I decided to analyze the risk factors and predict risk for each company. Based on the risk factor, I want to scrap the original resume and showcase only a part of it to the companies.

After conducting extensive research, I identified various factors that contribute to determining the legitimacy of a company. These factors were carefully documented and compiled into a survey. The survey was then distributed to individuals within My class and neighborhood as Shown in Figure1. It was encouraging to receive responses from over 45 participants, representing diverse age groups who regularly utilize LinkedIn for various purposes.

I meticulously analyzed the survey results, considering both quantitative and qualitative aspects. Through this comprehensive investigation, I identified 11 key features that were derived from the survey data. These features were incorporated into a dataset to train a Machine Learning model for predicting the legitimacy of companies on LinkedIn as shown in Figure 1.

Using these 11 features, I created a dataset. The dataset contains 11 features, and I have added a company name column to it. So, in total, the dataset now has 12 features. I utilized this dataset to create a model in ML that predicts whether the company is real or fake. I found the precision-recall accuracy and F1 score for all My implemented algorithms. When a feature input is fed to the

model, the model predicts whether the company is real or fake. I have assigned every company a risk label, using a scale of 0 to 5. A label of 0 indicates that the company carries no associated risk, while a label of 5 indicates that the company is considered to be at a high level of risk.

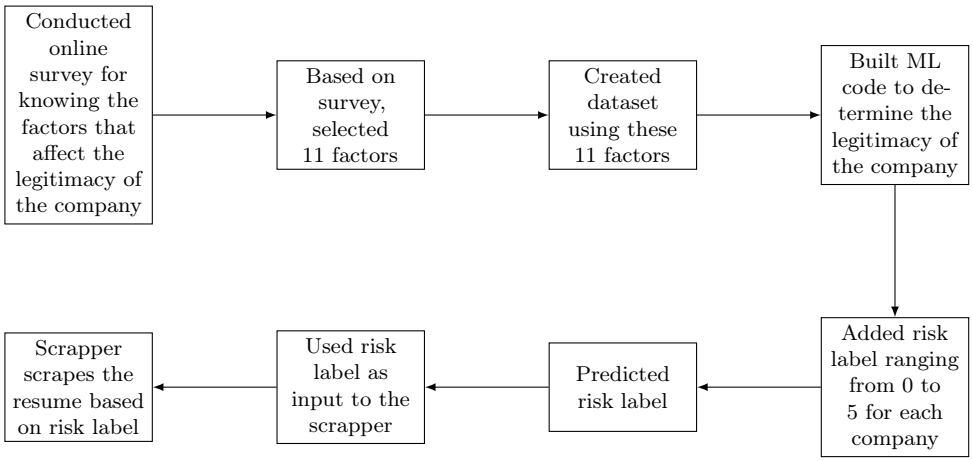


Fig. 1. High level flowchart of the project

2 Background and Related work

A paper was written by Pandya et al.[9] on "Detection of Deceptive Accounts Using Machine Learning Algorithm and Deep Neural Network" in 2020. In this paper, the authors tested the efficiency of machine learning techniques and deep neural networks in detecting fake accounts through their analysis of a dataset that includes both authentic and fake accounts. The results showed that deep neural networks, specifically convolutional neural networks (CNN), outperformed classical machine learning techniques in terms of accuracy.

A paper written by Sahoo et al.[11] on "Identification of Malicious Accounts on Facebook" in 2019 focused on classifying fraudulent and legitimate Facebook profiles using various machine learning algorithms. The authors tested classifiers such as decision trees, random forests, SVM, and KNN, and evaluated their effectiveness using metrics such as accuracy, precision, recall, and F1-score. Their methodology demonstrated reliable and precise identification of rogue accounts in a dataset comprising both account types.

The work of Khaled et al.[8] in 2018 on "Detecting Fake Accounts on Social Media" served as a reference. The authors gathered data from user profiles, posts, interactions, public blacklists, and user reports to identify distinct patterns of fake accounts. They employed machine learning algorithms such as decision trees,

random forest, logistic regression, and SVM for account classification, evaluating their performance using common evaluation criteria.

Adikari and Dutta's[1] paper in 2020 on "Identifying Fake Profiles in LinkedIn" addressed the problem of detecting fake profiles on LinkedIn. The authors proposed a data mining strategy to detect fraudulent profiles based on limited profile information. Their method achieved an accuracy of 87% and a True Negative Rate of 94%, demonstrating its effectiveness even with limited data. This study focused on detecting fake profiles rather than fake companies on LinkedIn.

After thoroughly examining these papers, I found that My project of utilizing a Machine Learning model[7] to detect the legitimacy of companies on LinkedIn incorporates a unique and novel aspect, the implementation of a scrapper. Previous research did not specifically focus on identifying fake companies on LinkedIn or integrating a scraper with Machine Learning. Therefore, My project marks the first-time implementation of Machine Learning and a scrapper to determine company legitimacy and extract relevant information from resumes based on risk labels. Unlike the paper by Adikari, My work specifically addresses the detection of fake companies on LinkedIn, making it a novel contribution in this domain.

3 Dataset

I have sent a survey link to both the class and the neighborhood. This survey aimed to gather information regarding the legitimacy of companies on LinkedIn. The survey included multiple questions that helped us identify the factors contributing to determining whether a company is real or fake. I received 45 responses from individuals of various age groups who use LinkedIn. After conducting a qualitative analysis of the survey data, I identified potential new features that could be added. Upon thorough examination of the survey results, I concluded that the top 11 features should be considered. Each feature will be assigned a value of 1 or 0. After considering all 11 features, I have decided to construct a dataset. I need a list of companies to create this dataset. The dataset now includes a column for company names. Therefore, the 12 features (I excluded the company name from My evaluation and counted it as datapoint ID) are as follows:

1. Company Name (Excluded in My machine learning analysis)
2. LinkedIn Followers > 1000
3. Profile Picture
4. Responsive Website
5. Staff Count > 1000
6. Staff Premium Accounts
7. Summary Section
8. Verified Address
9. Published Articles
10. Requests Sensitive Info
11. Legitimate Email
12. LinkedIn Recommendations

Fortunately, IBM and TCS have provided us with a comprehensive list of both fake and real companies. From this list, I have selected companies to include in My dataset, taking into account their classification as counterfeit or real. I have added a new column to the dataset specifically indicating whether a company is fake or real. This column, denoted as "Company Real or Fake," with values of 1 or 0, serves as the 13th feature in My dataset. After conducting a qualitative analysis by doing a survey, I extracted the top 5 most important factors among the 11 factors I have in My dataset as features. The list of top-5 features is shown in Table 1.

To ensure the accuracy of each feature value, I manually verified the data for a total of 100 companies. For each cell, I filled in the corresponding values for 50 real companies and 50 fake companies. Subsequently, utilizing the available real and fake company lists, I synthesized the data for the remaining cells based on the classification of the respective companies. As a result, the dataset now comprises 801 rows, encompassing the collected information.

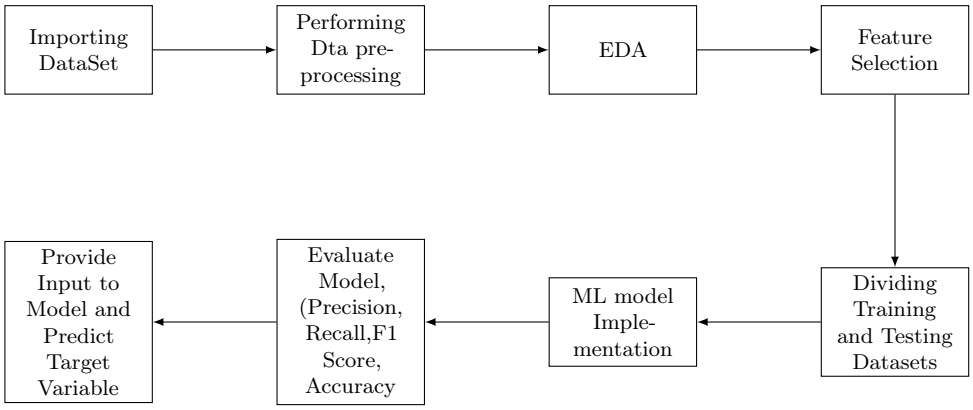


Fig. 2. Steps for finding out the legitimacy of the company

Table 1. List of Top-5 factors obtained from the dataset features.

Legitimate Email
Profile picture
Responsive website
verified Address
Request Sensitive Info

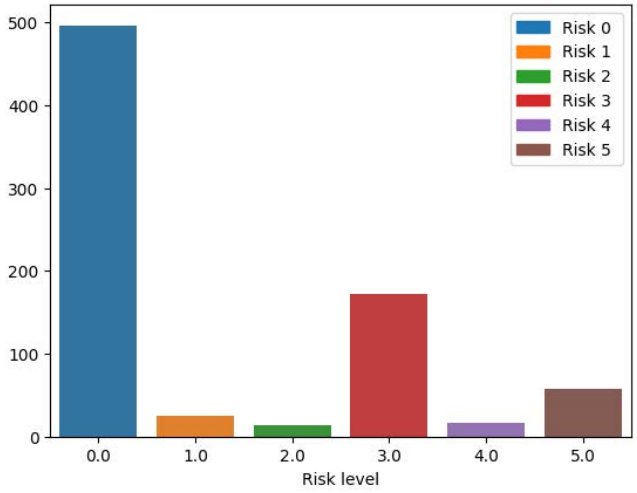


Fig. 3. The frequency of data points for each risk level class.

4 Design and Analysis

In this section, I explained the machine learning method I employed in my datasets. It covers all subsections, including Model selection, Imbalanced data, k-fold cross-validations, and experimental results.

Model Selection Model selection in machine learning is choosing the most appropriate algorithm or model for a given problem. It involves evaluating and comparing different models based on their performance and selecting the one that best meets the requirements and objectives of the task at hand. In this project, I applied three well-known applications, including Support Vector Machine, Decision Tree, and K-Nearest Neighbor(KNN) approaches to evaluate My model's performance. In case of model selection, I follow up the structured machine learning application to predict the risk level and company legitimacy (Real or Fake). Firstly I define the problem I are trying to solve and start to determine the requirement for My classification task.

Firstly, the initial step involves importing the dataset into either a Jupyter Notebook or Google Colab. The subsequent step entails verifying the presence of null values. As My dataset contained minimal null values, I eliminated rows with null values. The following step in data preprocessing involves examining for duplicate values. Since My dataset had no null values, no modifications were made. Since the majority of the data consists of numerical values, this aspect was satisfactory. However, the "Company Name" column contained company names represented by alphabets. Therefore, I employed Label encoding to convert the data in that column into a numerical form. The data must undergo preprocessing

and cleaning procedures to ensure that it is in an appropriate format for modeling purposes.

Following that, I proceeded with exploratory data analysis (EDA), wherein I analyzed the data by creating several graphs to gain further insights. Finally, for feature selection, I utilized a heatmap to determine the correlation between each feature in the dataset. I then performed a correlation analysis of all the features with the target feature, which is "Company Real or Fake." I sorted the features in descending order to identify the most highly correlated ones, as shown in Figure 5. However, since this analysis is experimental, relying solely on a few features with high correlation is not advisable. Consequently, I decided to consider all features when predicting the legitimacy of a company.

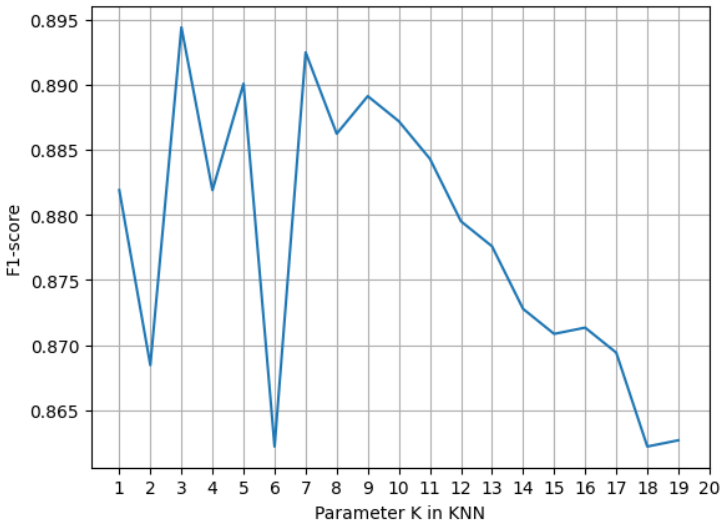


Fig. 4. The f1-score value for different k-value in KNN model.

Split the data: Divide the available data into training, validation, and test sets. The training set is used to train the models, and the test set is used to assess the selected model's final performance. In my experiments, I employed 30 percent of the datasets for the test set and used the rest to train the models.

Evaluation metrics are employed to evaluate and compare the performance of different models, and choosing evaluation metrics for machine learning is problem-dependent. In this project, we employed well-known classification metrics, including accuracy, precision, recall, and F1-score, to evaluate the model's performance on different models. The formula of Precision, Recall, and F1 score is presented in Equation1-3, respectively, where P denotes the precision, R denotes the Recall, and n_{tp}, n_{fp}, and n_{fn} indicate the numbers of true positives, false positives, and false negatives, respectively.

$$Precision = \frac{n_{tp}}{n_{fp} + n_{tp}} \tag{1}$$

$$Recall = \frac{n_{tp}}{n_{fn} + n_{tp}} \tag{2}$$

$$F1score = \frac{2|PxR|}{|P| + |R|} \tag{3}$$

In the case of training the model, I also applied the hyper-parameter optimization to find the best parameters in My employed model. For instance, I put the visualization of the hyper-parameter optimization on KNN model for different K, as shown in Figure4. This plot shows that the best k value for the KNN model is 3, which gives the highest F1 score compared to the k value.

These are the features of correlation.

```
In [19]: corr_matrix = generalData_df.corr()
         print(corr_matrix['Company Real or Fake(1 or 0)'].sort_values(ascending=False))

Company Real or Fake(1 or 0)       1.000000
LinkedIn Followers greater 1000    0.877671
Staff Count greater 1000           0.780602
Staff Premium Accounts             0.751040
LinkedIn Recommendations           0.728332
Published Articles                 0.716473
Verified Address                   0.692565
Requests Sensitive Info            0.284186
Summary Section                    0.249746
Legitimate Email                   0.110529
Responsive Website                -0.003343
Profile Picture                   -0.092269
Name: Company Real or Fake(1 or 0), dtype: float64
```

Fig. 5. Correlation with respect to company real or fake in descending order

To predict whether a company is real or fake, my task involves a binary classification, where the prediction is either a "yes" or "no" outcome. This implies that I am distinguishing between two classes. Therefore, for this particular problem, I opted for machine learning models that fall under the classification type. Conversely, since my prediction does not involve continuous values, I did not choose regression models.

In order to predict the risk level, I also applied multi-classification models to predict the risk level in My applications. Regarding the risk level generation, I used a regularization factor, which rewards the top 5 important factors against the other features in My dataset. To solve this, I start to generate a function. for handling the range of the generated value and also controlling the level of rewards to the most important factors, I determine the $\alpha = 1.19$, based on Equation4, I assume all the legitimate company Risk level 0, so I applied this function on the companies which are fake and assume $n = 5$ for the features(f_i) for the top-5

features. After getting the risk value from Equation4, I categorized the risk level based on Table2 to map the risk value to the proper risk label.

$$R = \begin{cases} 0, & \text{All legitimate companies.} \\ \sum_{n=1}^{11} \alpha^n f_i & \text{otherwise.} \end{cases} \tag{4}$$

To proceed with My analysis, I divided the dataset into training and testing sets as shown in Figure 2. I initially implemented the Decision Trees algorithm. This supervised machine-learning approach is used for both classification and regression problems. It constructs a model with a tree-like structure to make decisions based on input features.

The root node, representing the complete dataset, serves as the starting point in the decision tree. The algorithm then divides the data at each node into multiple classes or minimizes overall variance (in the case of regression) based on a chosen feature. Typically, criteria such as Gini impurity or information gain are utilized to determine the most suitable feature for division. For this model, I

Table 2. Risk level range for categorizing the risk value.

Range	Risk Level
[0,5]	5
(5,7]	4
(7,9]	3
(9,10]	2
(>10)	1

calculated metrics such as Precision, Recall, F1 score, and support. The accuracy achieved by this algorithm was 95.65%

Having trained the model, I am now able to provide 12 feature inputs to the model, which in turn predicts whether the company is real (0) or fake (1) as shown in Figure6 and Figure7

```
In [61]: features = np.array([[0,0,1,0,0,0,1,0,1,0,0,1]])
         # using inputs to predict the output
         prediction = ml.predict(features)
         print("Prediction: {}".format(prediction))

         Prediction: [0]
```

Fig. 6. Model Predicted the company is fake

```
In [62]: features = np.array([[0,0,1,1,0,1,1,0,1,0,0,1]])
         # using inputs to predict the output
         prediction = ml.predict(features)
         print("Prediction: {}".format(prediction))

         Prediction: [1]
```

Fig. 7. Model predicted the company is real

Imbalanced Dataset As I explained in the prior section, the frequency of each risk label is distributed unequally. The number of data points with a risk level of 0 is more than other risk labels, which causes the machine learning model to have a lower prediction accuracy. The frequency of the datapoint for each class of risk level is shown in Figure3. To tackle this problem, I applied an over-sampling approach, which synthesizes the data points in the classes that have a minority to the other risk label. The over-sampling method that I used is the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is an oversampling algorithm that addresses the issue of imbalanced classification datasets, where the representation of classification categories is unequal. The SMOTE technique generates synthetic data for the purpose of oversampling through the utilization of a k-nearest neighbor algorithm. The algorithm conducts an iterative process that traverses each observation belonging to the minority class. During this process, it identifies the k nearest neighbors of each observation and subsequently chooses a random set of neighbors for the generation process. The quantity of utilized neighbors is contingent upon the necessary level of oversampling. The SMOTE technique produces plausible synthetic instances from the underrepresented class, meaning that they are situated in the feature space in proximity to the pre-existing instances from the underrepresented class.

I applied the SMOTE, which increased the number of data points from 800 to 2796, which led to an increase in the number of data points to satisfy the model's robustness and completeness in the case of prediction accuracy. Based on the quantitative results which are shown in Figure 11, Although the SVM's accuracy in total is around 90 percent, the model's accuracy for class2 and class 4 is 0; it means that the model can not be trained for that classes and the main reason for this occurrence can be the low amount of labeled value for these classes. By applying the SMOTE techniques, the model's accuracy for class 2 and class 4 is increased to 85% and 75%, respectively. The quantitative results for the three models after and before the over-sampling approach is shown in Figure11-14 Moreover, a confusion matrix is also added to show the frequency of correct and wrong predictions for each class in the applied model, and all of the determined confusion matrices are also shown from Figure 8-10 efficiently.

4.1 K-fold cross validation

K-fold cross-validation is a method for analyzing the efficacy and generalizability of a machine-learning model. The given dataset must be divided into k equally sized subgroups, or "folds," where k is a fixed integer. The model is trained and tested k times, with each fold acting as a validation set once and being used to train the remaining folds. In this project, as I mentioned in prior sessions, My dataset encountered an imbalance challenge, which I solved using the SMOTE approach. I applied k-fold cross-validation to prove the model's performance and robustness on applied datasets. I also showed the results of the applied model by passing the k-fold cross-validation techniques to confirm the generalizability of the model and applied approaches. The quantitative results of the three models are shown in Table3.

Table 3. k-fold cross validation results for three employed models.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
KNN	0.7835	0.9042	0.8218	0.8956	0.8672
SVM	0.8724	0.8436	0.8722	0.8403	0.8521
Decision Tree	0.9144	0.9126	0.9058	0.9126	0.8941

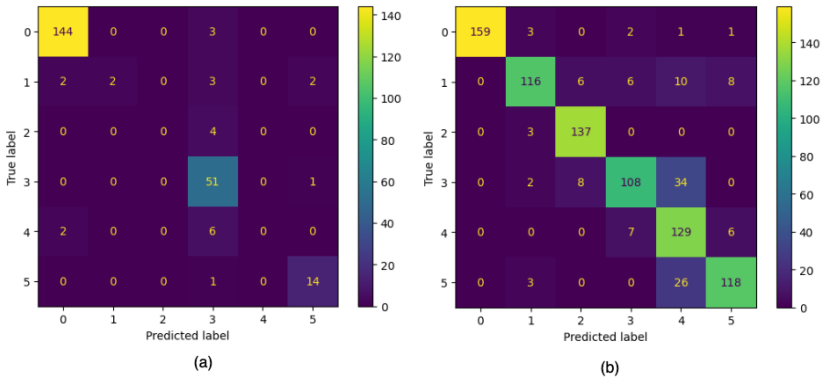


Fig. 8. (a) confusion matrix before over-sampling on SVM and (b) is the confusion matrix after SMOTE technique.

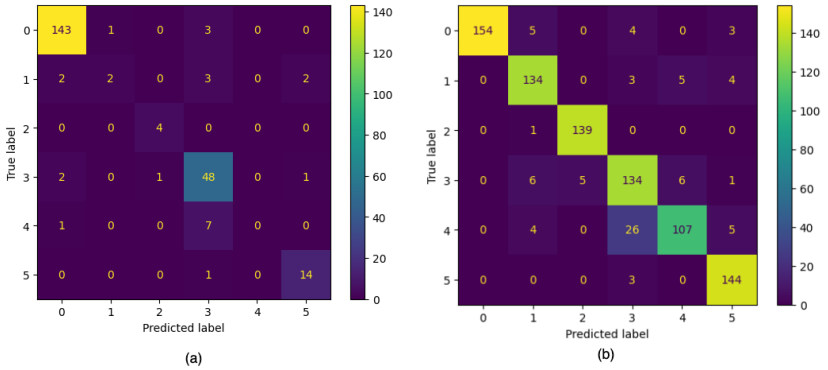


Fig. 9. (a) confusion matrix of Decision tree before over-sampling and (b) is the confusion matrix of Decision tree after SMOTE technique.

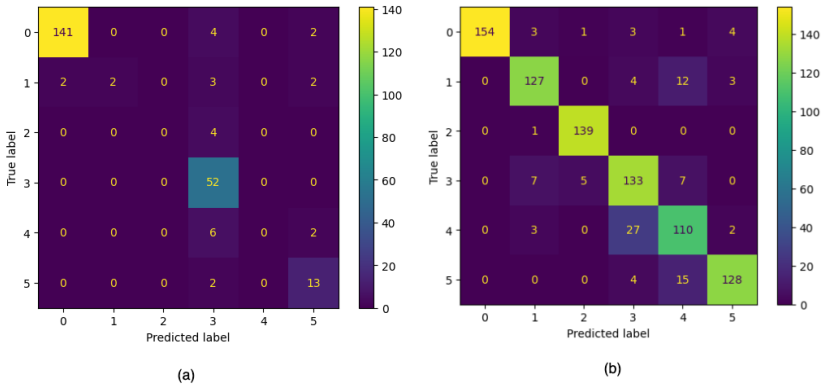


Fig. 10. (a) confusion matrix of KNN before over-sampling and (b) is the confusion matrix of KNN after SMOTE technique.

5 Risk Factor Analysis

The next step in the risk management process involves classifying the identified risk factors based on their potential impact and likelihood of occurrence. To ensure consistency and objectivity in the risk assessment process, a 0-5 scaling system is used to classify the risk factors. The 0-5 scaling system categorizes the risk factors into six levels based on their potential impact and likelihood of occurrence. The levels are shown in Table 4:

- **Level 0: No Risk** - Risk factors that are unlikely to occur or have a minimal impact if they do. No further action is required for these risk factors.
- **Level 1: Low Risk** - Risk factors that have a low likelihood of occurring or have minimal impact if they do. These risk factors can be monitored and managed through routine procedures.

	precision	recall	f1-score	support
class 0	0.97	0.98	0.98	147
class 1	1.00	0.22	0.36	9
class 2	0.00	0.00	0.00	4
class 3	0.75	0.98	0.85	52
class 4	0.00	0.00	0.00	8
class 5	0.82	0.93	0.87	15
accuracy			0.90	235
macro avg	0.59	0.52	0.51	235
weighted avg	0.87	0.90	0.87	235

Fig. 11. (a) The summary of the accuracy report for each class on the SVM model before the implementation of SMOTE.

	precision	recall	f1-score	support
class 0	1.00	0.96	0.98	166
class 1	0.91	0.79	0.85	146
class 2	0.91	0.98	0.94	140
class 3	0.88	0.71	0.79	152
class 4	0.65	0.91	0.75	142
class 5	0.89	0.80	0.84	147
accuracy			0.86	893
macro avg	0.87	0.86	0.86	893
weighted avg	0.88	0.86	0.86	893

Fig. 12. (a) The summary of the accuracy report for each class on the SVM model after the implementation of SMOTE.

- **Level 2: Moderate Risk** - Risk factors that have a moderate likelihood of occurring and could have a moderate impact if they do. These risk factors require careful monitoring and management to prevent or mitigate their effects.
- **Level 3: High Risk** - Risk factors that have a high likelihood of occurring and could have a significant impact if they do. These risk factors require immediate attention and proactive management to prevent or mitigate their effects.
- **Level 4: Very High Risk** - Risk factors that are almost certain to occur and could have severe consequences if they do. These risk factors require urgent action and special measures to prevent or mitigate their impact.
- **Level 5: Extreme Risk** - Risk factors that are virtually guaranteed to occur and could have catastrophic consequences if they do. These risk factors require immediate and decisive action to prevent or mitigate their impact.

Once I have classified the company as real or fake by using Machine Learning algorithms with risk levels 0-5, the next step is to determine the risk factors

	precision	recall	f1-score	support
class 0	0.97	0.97	0.97	147
class 1	0.67	0.22	0.33	9
class 2	0.80	1.00	0.89	4
class 3	0.77	0.92	0.84	52
class 4	0.00	0.00	0.00	8
class 5	0.82	0.93	0.87	15
accuracy			0.90	235
macro avg	0.67	0.68	0.65	235
weighted avg	0.87	0.90	0.88	235

(a)

	precision	recall	f1-score	support
class 0	1.00	0.93	0.96	166
class 1	0.89	0.92	0.91	146
class 2	0.97	0.99	0.98	140
class 3	0.79	0.88	0.83	152
class 4	0.91	0.75	0.82	142
class 5	0.92	0.98	0.95	147
accuracy			0.91	893
macro avg	0.91	0.91	0.91	893
weighted avg	0.91	0.91	0.91	893

(b)

Fig. 13. (a) The summary of the accuracy report for each class on the Decision Tree model before the implementation of SMOTE and (b) is the accuracy report of Decision Tree after the SMOTE implementation.

	precision	recall	f1-score	support
class 0	0.99	0.96	0.97	147
class 1	1.00	0.22	0.36	9
class 2	0.00	0.00	0.00	4
class 3	0.73	1.00	0.85	52
class 4	0.00	0.00	0.00	8
class 5	0.68	0.87	0.76	15
accuracy			0.89	235
macro avg	0.57	0.51	0.49	235
weighted avg	0.86	0.89	0.86	235

(a)

	precision	recall	f1-score	support
class 0	1.00	0.93	0.96	166
class 1	0.90	0.87	0.89	146
class 2	0.96	0.99	0.98	140
class 3	0.78	0.88	0.82	152
class 4	0.76	0.77	0.77	142
class 5	0.93	0.87	0.90	147
accuracy			0.89	893
macro avg	0.89	0.89	0.89	893
weighted avg	0.89	0.89	0.89	893

(b)

Fig. 14. (a) The summary of the accuracy report for each class on the KNN model before the implementation of SMOTE and (b) is the accuracy report of KNN after the SMOTE implementation.

associated with sensitive information and privacy factors in resumes. This allows us to identify which fields are required and which should be scrapped based on their associated risk levels. By classifying these risk factors, I can ensure that any sensitive information included on a resume is handled appropriately and that job seekers are protected from potential misuse of information. Ultimately, this process helps to prevent the leaking any information from an individual to fake companies and provides better information to the real companies.

I have used this scaling system so that organizations can prioritize their risk management efforts and allocate resources effectively to manage the risks that pose the greatest threat to their operations.

But in real-time scenarios, it is important to note that the risk classification should be reviewed regularly and updated as needed to ensure that the risk management strategy remains effective and relevant on a regular basis.

There are a few privacy-related things that can be mentioned in a resume. So again I am classifying the things into different categories:

Levels	Risk	Description
Level 0	No Risk	Have a minimal impact. No further action is required.
Level 1	Low Risk	Low likelihood of occurring. It can be monitored and managed through routine procedures.
Level 2	Moderate Risk	A moderate impact. Careful monitoring and management are required to prevent or mitigate their impact.
Level 3	High Risk	Have a significant impact. Require immediate attention and proactive management to prevent or mitigate their impact.
Level 4	Very High Risk	Have severe consequences. Urgent action and special measures are required to prevent or mitigate their impact.
Level 5	Extreme Risk	Have catastrophic consequences. Require immediate and decisive action to prevent or mitigate their impact.

Table 4. Classifying risk levels.

5.1 Contact Information:

Resumes typically include the Applicant's name, phone number, email address, and home address. While this information is necessary for potential employers to contact the applicant, it's important to consider the potential privacy implications and divide the risk factors accordingly. So usually in the contact information I have following categories.

Applicant's Name:- No risk Including one's name on a resume is generally considered a No-risk activity. In fact, it is a necessary and expected part of any resume, as it is the primary identifier for the job applicant. Including one's name on a resume is typically required to allow employers to match the resume with the job application and to contact the job seeker for further information or an interview. It is also useful for employers to keep track of multiple applicants throughout the hiring process. Even if a company is fake, it could have done nothing by getting a name because there can be multiple persons with similar names. In some cases, job seekers may choose to use only their first name or a nickname on their resume to avoid any potential biases or discrimination based on their gender, race, or ethnicity. This approach may not be practical for all job seekers and may depend on the industry and job being applied for.

Phone Number:-Very High Risk The Phone Number is typically considered low-risk in a resume. I have classified it because a phone number is a common and essential piece of contact information that is often provided voluntarily by the job seeker. While it is still important to protect this information from potential

misuse or data breaches, it is not considered to be as sensitive or private as other information like social security numbers, financial information, or home addresses. But nowadays, the mobile phone has become a way of life for everyone.So if they hack the phone, then every sensitive information one can get, especially when it comes to female candidates, is very high risk. Also, for example, the job seeker may receive unsolicited or threatening calls from telemarketers, scammers, or other unwanted callers. So, I am classifying this as a very high risk.

Email Address:- High risk The Email Address is typically considered to be at a high-risk level in a resume. This is because an email address is a common and necessary piece of contact information that is often provided voluntarily by the job seeker. The level of risk associated with an email address on a resume can vary depending on various factors. One such factor is the industry in which the job seeker works. For example, suppose someone works in a highly sensitive industry like defense or intelligence. In that case, their email address may be considered to have a higher risk because it could be a potential target for hackers or cyber attackers seeking to gain sensitive information. Another factor that can influence the level of risk associated with an email address is the job seeker's seniority level. For instance, if someone is a high-ranking executive or a person in a position of authority, their email address could be considered to have a higher risk because it may be more valuable to cyber attackers seeking to gain access to confidential information. If I consider a fresher who is applying to the job role it is regarded as a low level of risk, but on LinkedIn, not only freshers, there could even be experienced candidates, and more frequently, if they know the email addresses, they get spam emails, but as per privacy challenges if one hacks the email all the sensitive information gets revealed. And also, I can see if I want to sign up for an application, I can do it through My email, which is an identification for the apps. So, overall I are classifying that as High-risk.

Home Address:-Extreme risk Generally speaking, the Home Address is considered to be an extreme risk. This is because sharing one's home address on a resume can potentially expose them to various risks, including identity theft, burglary, or physical harm.[2] For instance, if a job seeker lives in an area with high crime rates or if their profession involves working with sensitive information, their home address could be considered to have a higher risk. Additionally, job seekers who are concerned about their privacy or safety may choose not to include their home address on their resume altogether, opting instead to provide their city, state, or region. Hence, the Home address I am classifying as high risk.

5.2 Education History: Moderate risk

Resumes typically contain information about the candidate's educational background, including the name of the institution, coursework, degree obtained with CGPA, and graduation date. As candidates usually include their entire education history from schooling to graduate levels, I consider it as one entity.[5] Education

History is usually categorized as a moderate risk item in the order of risks for information on a resume. This is because most information in education history is already in the public domain, such as the name of the institution, degree earned, and graduation date. However, there are some situations where an individual's education history may pose a higher risk. For example, if the institution attended is known for controversial or sensitive issues, such as a religious or political institution, their education history may be considered to have a moderate risk. Additionally, if a job seeker's education history includes sensitive information such as transcripts or academic records, it could be deemed to have a higher risk. Overall, Education History is generally considered to have a moderate risk.

5.3 Projects: Moderate risk

Here in projects, I will classify them into 2 ways.1 Academic Projects 2.Experienced Projects In terms of the risks for information on a resume, academic projects are typically considered to be low-no-risk items. This is because academic projects are often completed as part of a school curriculum and are not generally considered to be sensitive or confidential information. Additionally, the focus is typically on the project itself rather than personal information about the student or job seeker.

But if I consider them Experienced Projects, they involve highly confidential matters, such as clients' names and the work they have done, which is highly controversial and a privacy challenge. So, overall, I can consider the Projects to be "Moderate risk."

5.4 Work Experience:- Moderate risk

In the order of risks for information on a resume, work experience is typically considered to be a low-risk item. This is because work experience generally contains information that is already in the public domain, such as job titles, responsibilities, and duration of employment. Additionally, if a job seeker's work experience includes confidential information or trade secrets, it could be considered to have a higher risk. However, if an individual can inquire about a job seeker's details from the company if someone knows the company, it can pose a moderate risk.[3] Therefore, I classify this scenario as moderate risk.

5.5 Publications:- Moderate risk

In the order of risks for information on a resume, Publications are typically considered to be a no-risk item. This is because publications are usually intended to be shared with a wider audience and are already in the public domain.

But as I got input from my friends during the presentation, it is considered high risk because if someone gives the doi of the paper and one clicks the doi, then all the personal information gets leaked, so overall, publications are considered moderate risk.

5.6 Achievements:- Low risk

In the order of risks for information on a resume, Achievements are typically considered to be low-risk items. This is because achievements are often related to public recognition or awards received by the candidate and do not usually reveal any sensitive or personal information. However, there may be some cases where achievements are considered to be more risky. For example, if the achievement is related to a sensitive or controversial topic, it could be classified as a moderate risk. Additionally, if the achievement reveals personal or sensitive information, such as membership in a particular group or organization, it could be considered a moderate or high risk. Hence, I generally classified Achievements as a low-risk item on a resume.

5.7 Personal Information:

Applicants should avoid including any personal information on their resume that could be used to discriminate against them, such as age, gender, race, religion, or marital status. In some countries, this type of information is illegal for employers to ask for during the hiring process. Usually, many people in any country don't specify the personal information in their resume they usually do it in CV. But if any person mentions this I want to classify the risk factors and scrap the resume accordingly.

Age:- Moderate risk Age is generally considered a no-risk item on a resume, as it is not directly related to a person's qualifications for a job. In fact, including age on a resume is often discouraged due to potential age discrimination. However, in some cases, indicating a specific age range (e.g., "recent college graduate" or "10+ years of experience") may be necessary or preferred by the employer here. I am not concerned with any discrimination, but I am focussing on privacy [4] regulations if one gets the name and age of a person, there could be many places with the database having name and age, so I would typically be considered a moderate risk item.

Gender:- moderate risk In general, disclosing gender in a resume is considered a low-risk item. This is because gender is often readily apparent from the applicant's name, and it is not typically considered sensitive information. However, as per the privacy challenge, this is sensitive information, so I would classify it as a moderate risk.

Race:-Moderate risk: Mentioning one's race or ethnicity on a resume can be a highly sensitive issue and is generally not recommended. As such, I would classify it as a high-risk activity. While there may be instances where providing race or ethnicity information on a resume may be relevant, such as when applying for jobs where diversity is a critical factor, in most cases, it is not necessary or advisable. Including such information on a resume can create the perception of

bias or discrimination, even if unintentional. Additionally, in many countries, including the United States, it is illegal for employers to discriminate against job applicants on the basis of race or ethnicity. Therefore, including this information on a resume may not provide any real advantage to the job seeker. Overall, apart from the discrimination, a privacy challenge would classify a race or ethnicity as a Moderate risk

Religion:- High-risk activity Religion is a highly personal and sensitive topic, and including it on a resume can potentially lead to discrimination or bias. Therefore, I would classify mentioning religion on a resume as a high-risk activity. On the one hand, some job seekers may choose to include their religion on their resume if it is relevant to the position or the company's mission, values, or culture. For example, a job seeker applying to a religious organization or a faith-based non-profit may choose to highlight their religious affiliation or involvement as a qualification for the job. On the other hand, some employers may view an applicant's religion as irrelevant to the job and potentially discriminatory. In some countries, including the United States, employers are prohibited by law from discriminating against job applicants based on their religion. However, even if it's not illegal to ask about religion in a particular country, it can still create the perception of bias or prejudice.

Marital Status:- Moderate risk activity In general, marital status is considered a low-risk item on a resume. This is because it is not typically relevant to a person's qualifications for a job, and it is also protected information in many countries. However, in some situations, marital status could be considered a moderate risk item. For example, if the job seeker is applying for a position that requires extensive travel or relocation, the employer may be interested in the candidate's marital status. Additionally, in some cultures or industries, marital status may be considered relevant or important information for identifying a person. So, as a privacy challenge, I consider it a moderate-risk activity.

But mostly personal information has nothing to do with the company unless discrimination, but as a privacy challenge, I would classify that to be very important and considered as high risk and moderate risks

5.8 Categorizing the risk factors

How can I effectively extract information from a resume while taking into account the risk factor levels ranging from 0 to 5 for each element in the resume? It is shown in Table5.

- **Risk Factor Level 0:** The resume should be provided as is, without any changes or field scrapping, as the data contained in the resume has minimal impact and no action is required.

- **Risk Factor Level 1:** The resume should be provided with all features except for those ranked at Risk Level 5. This means excluding the home address while keeping all other fields intact.
- **Risk Factor Level 2:** The resume should be provided with all features except for those ranked at Risk Level 5 and Level 4. This entails excluding the home address and phone number while retaining all other fields.
- **Risk Factor Level 3:** The resume should be provided with all features except for those ranked at Risk Level 5, Level 4, and Level 3. This involves excluding the home address, phone number, email address, and religion or ethnicity while keeping all other fields.
- **Risk Factor Level 4:** The resume should be provided with all features except for those ranked at Risk Level 5, Level 4, Level 3, and Level 2. This means excluding the home address, phone number, email address, religion or ethnicity, education history, work experience, age, gender, race, and marital status while retaining all other fields.
- **Risk Factor Level 5:** The resume should be provided with all features except for those ranked at Risk Level 5, Level 4, Level 3, Level 2, and Level 1. This entails excluding the home address, phone number, email address, religion or ethnicity, education history, work experience, age, gender, race, marital status, achievements, and profile picture while keeping all other fields intact.

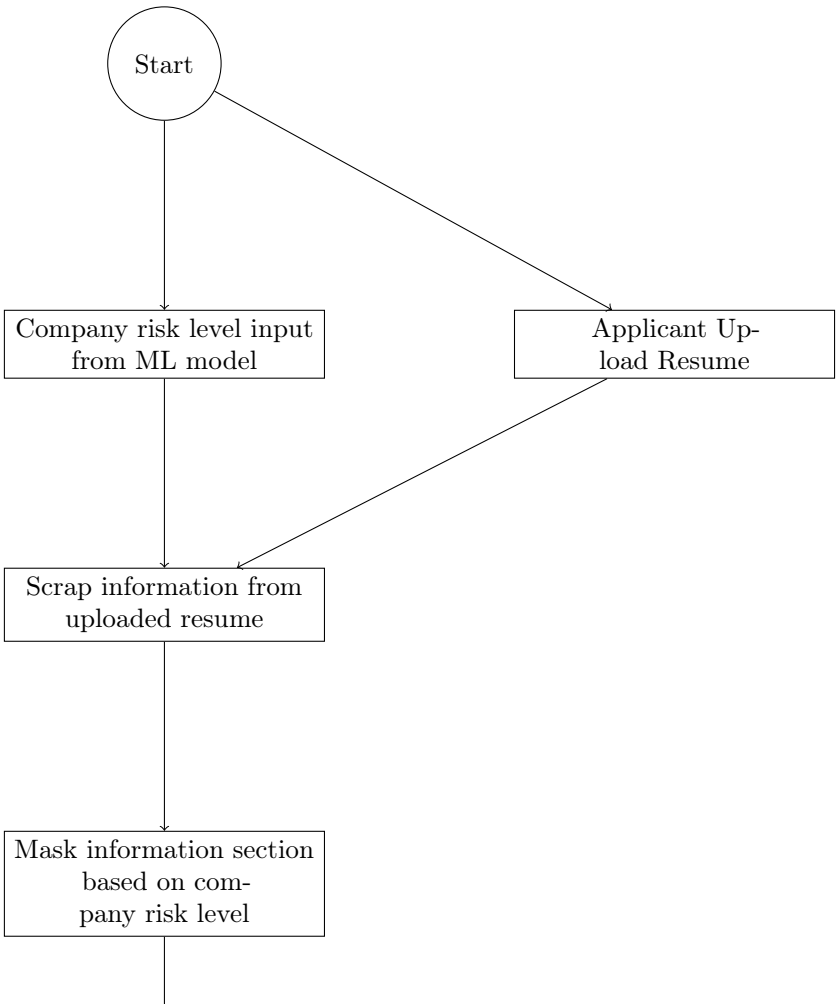
6 Scraper

The scraper is written in Python. It is implemented to help LinkedIn users protect their private information from their resumes when they apply for a job. Users may want to share information with a company, depending on the company's legitimacy. If a company does not seem to be so legit, users may want to share less information from their resume, and vice versa. This is a benefit for users when they are trying to apply for jobs at many companies. With Scraper, users will not need to modify their resumes for each of the companies they are applying to. The scraper will take the user's original resume and then generate a new version of the resume for the specific company that the user used for, depending on their risk factor score. The new version of the resume will have the same information as the original resume, but hiding some information depends on company risk factor criteria.

For the scraper implementation, I assume that users' resumes should be a single template and format to lessen the complexity of coding.

Levels	**Risk**	**Description**
Level 0	No Risk	No changes or scrapping required.
Level 1	Low Risk	All features except for those ranked at Risk level 5 (excluding the home address).
Level 2	Moderate Risk	All features except for those ranked at Risk Level 5 and Level 4 (excluding the home address and phone number).
Level 3	High Risk	All features except for those ranked at Risk Level 5, Level 4, and Level 3 (excluding the home address, phone number, email address, and religion or ethnicity).
Level 4	Very High Risk	All features except for those ranked at Risk Level 5, Level 4, Level 3, and Level 2 (excluding the home address, phone number, religion or ethnicity, education history, work experience, email address, publications, age, gender, race, and marital status).
Level 5	Extreme Risk	All features except for those ranked at Risk Level 5, Level 4, Level 3, Level 2, and Level 1 (excluding the home address, phone number, religion or ethnicity, education history, work experience, email address, publications, age, gender, race, and marital status).

Table 5. Categorizing risk factors and determining what needs to be scrapped in a resume .



6.1 Scraping Process

Upon running, the scraper receives two inputs: the applicant's uploaded resume and the risk level received from the ML model that predicts company risk. The first process after getting inputs is for the scraper to start to scrape the resume. The resume uploaded has a `docx` format and is converted to `txt` format. It is then read and parsed using a spaCy entity. Each section in the resume, including name, address, phone number, email, education, project, experience, skills, and certificate, is located and stored separately. For each of these sections (excluding name, skills, and certificate), a duplicate is created, and sensitive information is masked.

- **Address:** A duplicate is created with the covered number and street name. I.e., "8465 Foundry St, Savage MD 20763" is masked as "**********, Savage MD 20763".
- **Phone:** A duplicate is created with covered last four digits. I.e., (240)-381-4696 is masked as (240)-381-****.
- **Email:** A duplicate is created with the email being masked until the "@" character is found. I.e., atran12345@umbc.edu is masked as **********@umbc.edu.
- **Education:** For each school in the education list, the name of the University/College is covered. I.e., "University of Maryland, Baltimore County" is masked as "University of **********." The remaining items are kept as is, including the degree, GPA, honors, etc.
- **Projects:** For each project in the projects list, the time and location where the project was implemented are covered. I.e., "TCP Socket - Multiusers Concurrent Client/Servers – UMBC Fall 2022" is masked as "TCP Socket **********."
- **Experience:** For each job history in the experiences list, the company name, location, and time associated with the job are covered. I.e., "Math Tutor - Howard Community College, Columbia, MD August 2016 - May 2018" is masked as "Math Tutor **********."

After creating all duplicate sections with covered information, the scraper starts to generate a resume based on the risk level received from the ML model. Each risk level is associated with different private information that needs to be covered, and a resume is generated accordingly.

- **Risk 0:** Generate a resume with all original information.
- **Risk 1:** Generate a resume with the address covered.
- **Risk 2:** Generate a resume with the address and phone covered.
- **Risk 3:** Generate a resume with the address, phone, and email covered.
- **Risk 4:** Generate a resume with the address, phone, email, and education covered.
- **Risk 5:** Generate a resume with the address, phone, email, education, projects, and experience covered.

For resumes that have covered information, the generated resume has an ending statement: "Applicant's preferences are protecting this resume." To notify

this, the applicant has to opt-in to have data privacy protection. The resume that they have received from the applicant is not the original resume. They could potentially contact applicants through Linked In or another type of contact information shown in the resume if they are available to request an original resume.

6.2 Scraper Implementation

Resume processing . Scraper reads resumes in the `docx` format and utilizes the `docx2txt` library to convert the `docx` documents into text format. The library's `load` function is then employed to create an entity capable of reading the text document. Specifically, the entity is trained on written web text and encompasses vocabulary, syntax, and entities in English, using the `en_core_web_sm` pipeline.

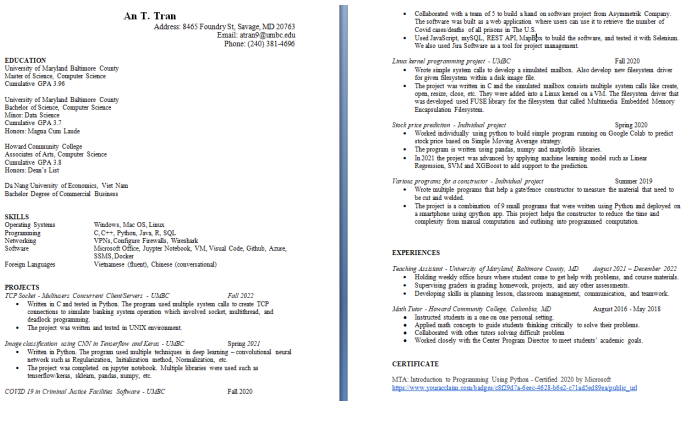
Information locating After successfully reading the resume, the scraper starts to locate information sections. This process uses three techniques. **Spacy library:** An NLP entity using the Spacy library is used to locate names and emails [10]. Other than name and email, spacy does not work efficiently over other types of information. This is because the dictionary used for the Spacy model trained is not specified for scarping resumes, but only general webs using English[6]. **Regular Expression:** The address and phone number are located using regular expression techniques. Two predefined regex formulas are used to locate the address and phone number in the resume. For each piece of information, the scraper searches through the whole resume to match everything that matches as a predefined regex and returns accordingly. **Key words:** The last technique that is implemented is keyword searching. A set of keywords, "Education, Skills, Projects, Experiences, Certificate," is predefined to be used to search for each information section in the resume. For each section that is associated with a keyword, scarper will search throughout the resume to locate that keyword and store the location information. The search is done using regular expression search. Then Scarper starts another search from the location found to the end of Scarper. The search stops when another keyword is detected or ends the resume and stores the section stopping location information. I.e., for the "Projects" keyword, scarper will search through the resume until "Projects" is located. Then Scarper starts another search from the "Projects" location to the end of the resume until any keyword in "Education, Skills, Experiences, Certificate" is located. The location value has the format [start, stop], and a section will be extracted from the resume from this return location information retrieved from the searching process.

Information masking After successfully locating all information sections, the scraper starts to mask all private information, including address, phone number, email, education, projects, and experiences. - Email masking: Scraper finds the "@" location, then replaces every character before this location with a "*" character. Phone number masking: The scraper removes the last four digits from a phone number and adds "****" at the end. - Address masking: Scraper finds the first

comma location in the address section, then replaces every character before this location with a "*" character. - Education masking: Scraper searches each line in the education section. A line contains the word "University" or "College," meaning that that line has the name of the school. The scraper will replace every character in that line, excluding "University of" or "College of" with a "*" character. Experience masking: The scraper searches each line in the experience section. If a line contains a "-" character, meaning that it has a job title, company name, location, and time separated by a "-" character, the scraper will replace every character in that line, excluding the job title (before the "-" character), with a "*" character. - Project masking: This process is like the process of experience masking, in which the scraper will replace every character in the line containing the project name with the "*" character, excluding the project name.

Resume generator: After successfully masking all information sections, Scarper starts to generate a resume base on company risk. For each risk level, corresponding information sections are sent to the generator function to create a resume accordingly. Scraper uses the docx library to generate docx documents for new resumes. The new resume has a name at the top of the document as a heading line. Follow is a paragraph generated by the docx entity that contains the address, phone, and email. For the education, skills, projects, experiences, and certificates sections, each is a separate paragraph with the section's name at the beginning as a heading line. For a company with a risk level greater than 0, the scraper adds a sentence, "This resume be being protected by applicant's preferences," at the end of the document. Finally, the scraper saves the new resume name with the file name as the company risk level.

The original resume is shown in figure 15 . If the risk level is 01, the scraped resume is shown in figure 16, and if the risk level is 05, the scraped resume is shown in figure 17.



An T. Tran
Address: 8465 Foundry St, Savage, MD 20763
Email: atran5@umbc.edu
Phone: (240) 381-4696

EDUCATION

University of Maryland Baltimore County
Master of Science, Computer Science
Cumulative GPA 3.96

University of Maryland Baltimore County
Bachelor of Science, Computer Science
Minor: Data Science
Cumulative GPA 3.7
Honors: Magna Cum Laude

Howard Community College
Associates of Arts, Computer Science
Cumulative GPA 3.8
Honors: Dean's List

Da Nang University of Economics, Viet Nam
Bachelor Degree of Commercial Business

SKILLS

Operating Systems: Windows, Mac OS, Linux
Programming: C, C++, Python, Java, R, SQL
Networking: VPNs, Configure Firewalls, Wireshark
Software: Microsoft Office, Jupyter Notebook, VM, Visual Code, Github, Azure, SSMS, Docker
Foreign Languages: Vietnamese (fluent), Chinese (conversational)

PROJECTS

TCP Socket - Multiusers Concurrent Client/Servers - UMBC *Fall 2022*
- Written in C and tested in Python. The program used multiple system calls to create TCP connections to simulate banking system operation which involved socket, multithread, and deadlock programming.
- The project was written and tested in UNIX environment.

Image classification using CNN in Tensorflow and Keras - UMBC Spring 2021
- Written in Python. The program used multiple techniques in deep learning – convolutional neural network such as Regularization, Initialization method, Normalization, etc.
- The project was completed on jupyter notebook. Multiple libraries were used such as tensorflow/keras, sklearn, pandas, numpy, etc.

COVID 19 in Criminal Justice Facilities Software - UMBC Fall 2020
- Collaborated with a team of 5 to build a hand on software project from Asymmetrik Company. The software was built as a web application where users can use it to retrieve the number of Covid cases/deaths of all prisons in The U.S.
- Used JavaScript, mySQL, REST API, Mapbox to build the software, and tested it with Selenium. We also used Jira Software as a tool for project management.

Linux kernel programming project - UMBC Fall 2020
- Wrote simple system calls to develop a simulated mailbox. Also develop new filesystem driver for given filesystem within a disk image file.
- The project was written in C and the simulated mailbox consists multiple system calls like create, open, rmsite, close, etc. They were added into a Linux kernel on a VM. The filesystem driver that was developed used FUSE library for the filesystem that called Multimedia Embedded Memory Encapsulation Filesystem.

Stock price prediction - Individual project Spring 2020
- Worked individually using python to build simple program running on Google Colab to predict stock price based on Simple Moving Average strategy.
- The program is written using pandas, numpy and matplotlib libraries.
- In 2021 the project was advanced by applying machine learning model such as Linear Regression, SVM and XGBoost to add support to the prediction.

Various programs for a constructor - Individual project Summer 2019
- Wrote multiple programs that help a gate/fence constructor to measure the material that need to be cut and welded.
- The project is a combination of 9 small programs that were written using Python and deployed on a smartphone using qpython app. This project helps the constructor to reduce the time and complexity from manual computation and outlining into programmed computation.

EXPERIENCES

Teaching Assistant - University of Maryland, Baltimore County, MD *August 2021 – December 2022*
- Holding weekly office hours where student come to get help with problems, and course materials.
- Supervising graders in grading homework, projects, and any other assessments.
- Developing skills in planning lesson, classroom management, communication, and teamwork.

Math Tutor - Howard Community College, Columbia, MD August 2016 - May 2018
- Instructed students in a one-on one personal setting.
- Applied math concepts to guide students thinking critically to solve their problems.
- Collaborated with other tutors solving difficult problem.
- Worked closely with the Center Program Director to meet students' academic goals.

CERTIFICATE

MTA: Introduction to Programming Using Python - Certified 2020 by Microsoft
https://www.youracclaim.com/badges/c8f2bd7a-6aec-4628-b6a2-c71a45ed189a/public_url

Fig. 15. Original Resume

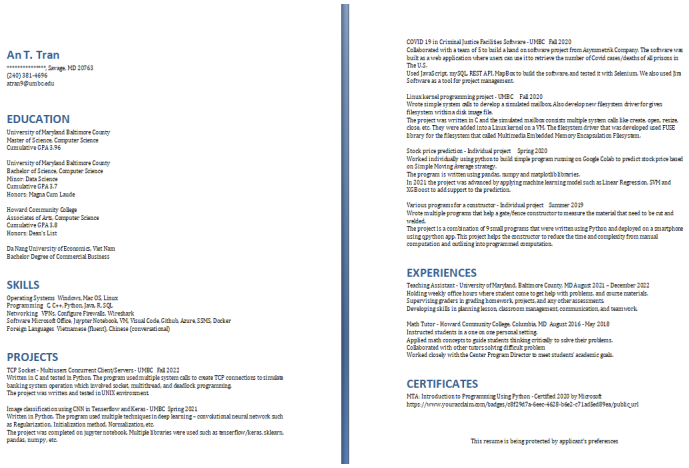


Fig. 16. Resume after being scraped for risk level 1

7 Discussion

In this project, I collected a list of fake and legitimate companies and stored it as a dataset; I applied qualitative and quantitative analysis and extracted the important features among the determined features. However, my dataset is not a real-world dataset, and my analysis can be advanced if I acquire my results based on the real-world dataset. I also conducted extensive analyses based on machine learning applications. I applied binary classification and multi-classification models to predict the company's legitimacy and risk level. Quantitative results also check the model's generalizability by employing k-fold cross-validation. So, I can figure out that the models are robust in the prediction of the risk level and the company's legitimacy.

Scraper covers private information in a way that is less likely to reduce an applicant's chance of being recruited. The scraper is doing this process not to cover information related to applicant capability. For example, for the experience section, the scraper only covers the company name, location, and job time. Still, it does not cover the experience that the applicant gains for that specific position. This helps the applicant remain attractive to the recruiter. Also, as another example, for the education section, the scraper only covers the school name. Still, it does not cover other information, such as the applicant's major, GPA, or any other achievement that the applicant has in that school. All of these were done so that applicants would still be able to show all their capabilities/skill sets to a recruiter for the position that they are trying to apply for.

Deploying resume scraper helps LinkedIn users protect their private information from risky companies. However, there is a utility trade-off when a user opts to apply to a company with a resume being scraped.

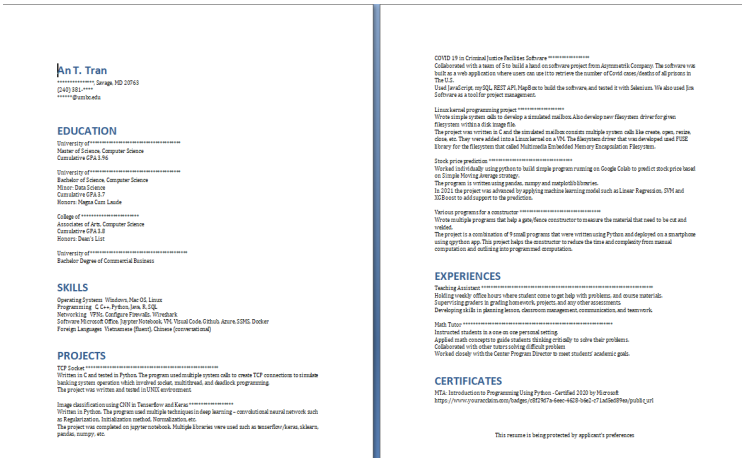


Fig. 17. Resume after being scraped for risk level 5

Applicant lost: When the user opts to use Scraper and apply to a company that is being detected as a high-risk company, the applicant will use a new protected resume with information covered to apply to the company. Also, the machine learning models have an accuracy of around 90%. Hence, there is a chance that the company that the applicant is trying to apply to is legitimate but is classified as a risky company. With a resume that has information covered, the applicant may lose the company's interest in the applicant's application. For instance, if another application has the same skill sets, experiences, history, and education as an application with a resume being masked, the recruiter is likely to choose the applicant that has an unprotected resume in which they can easily locate the applicant's contact information. Therefore, in this case, applicant loses their chance to get hired even if they meet the job's requirements.

Recruiter lost: Utility loss happens on the recruiter's side when there is a very good candidate who meets everything that the company needs for the hiring position. Still, the applicant opted to use the resume scraper. In addition, the company is legit, but it was accidentally detected as a risky company by the machine learning model. This led to the problem that the resume has covered contact information that make it difficult for the recruiter to contact this good candidate. On the other hand, if this also applies to another company that is not detected as a risky company, the other company may easily reach out to the applicant earlier and hire them. Hence, the legitimate company loses its chance to recruit a good applicant who meets all the requirements for the position that is missing.

At this stage of implementation, this resume scraper can only scrape resumes that are docx-type documents and have a similar format/structure. For future enhancement, I would need to exploit different NLP frameworks that suit best for

resume scraping so that it could scrape resumes with all types of structure. The predefined regex would also need to be improved so that it can detect different phone number formats and addresses. For instance, currently, the scraper is only able to detect addresses within the U.S.; however, it would need to be able to detect addresses for applicants who live outside of the country and have different address formats. Lastly, the set of keywords needs to be expanded so that it can detect information sections with various headlines. In addition, a variant of the word also needs to be considered. For instance, "Education" needs to be extended to include "EDUCATION, Education, education ...". The expanded list may get long, but this could easily be done by just adding more words to the current keyword list.

8 Conclusion

In conclusion, the study sought to address the critical problem of determining the legitimacy of organizations or companies on LinkedIn and projecting their appropriate risk classifications. A robust model was constructed to accurately classify whether a company is legitimate or deceptive by applying Machine Learning techniques, specifically the Decision Trees algorithm.

Additionally, the project's implementation of a scrapper component was a key component. Considering the risk labels projected for each organization, this scrapper was created to extract essential information from resumes. Well-defined policies were created to assure consistency and accuracy in the extraction process. These policies gave instructions on how to assign risk factors to particular resume items and choose which fields needed to be extracted and assessed.

An extensive study was done to identify the elements contributing to the company's legitimacy. This included looking through various resources, analyzing best practices in the business, and taking into account professional viewpoints. A thorough survey was created, comprising thoughtfully formulated questions that specifically targeted the factors related to the legitimacy of companies. The project team's class and neighborhood received the study, and an amazing response rate of over 45 participants was obtained. People from different age groups who actively use LinkedIn for various reasons make up this diverse group. Their insightful suggestions and contributions were thoroughly examined from both quantitative and qualitative angles.

The project team used rigorous quantitative analysis to identify 11 major features from the survey findings. These features, which included a wide range of properties and indicators, were then incorporated into the dataset. This dataset was used to train and fine-tune the Machine Learning model, allowing for accurate predictions about company legitimacy within the LinkedIn platform.

Overall, this project is a well-executed effort to address the important problem of knowing the legitimacy of companies on LinkedIn. Advanced Machine Learning methods combined with Scraper, data analysis, and policy creation have produced insightful findings and workable solutions for identifying and reducing risk in the

professional networking sector. The methodology and conclusions of this research have the potential to greatly improve the platform's overall trust and security.

9 Future Scope:-

1. The project has mainly two parts. I developed a machine learning algorithm in the first part of the project to determine whether a company is real or fake. I find out the risks associated with each company with the help of a risk label, and this risk label, which is predicted, is then inputted into the scrapper. The second part of the project is to develop a scraper. I feed the risk labels from the first portion into the scraper, which scrapes resumes depending on the risk labels. This enables us to extract essential information from the resumes for further research. I need a user-friendly interface that integrates both functionalities to allow smooth coordination between these two elements. I require a front-end system that allows users to interact with the project smoothly, performing functions from a single interface. This interface will make it easier to coordinate and run both portions of the project efficiently.
2. In terms of the project fields within users' specific LinkedIn profiles, they are assigned a risk level of 2, indicating a moderate level of risk. However, it is important to consider the possibility of certain projects carrying a higher risk. This is particularly relevant when a LinkedIn user showcases both academic and company-related projects they have worked on.
   The policy I have established indicates that all projects should be assigned a risk level of 2 without distinguishing between academic and company-related projects. However, this approach raises concerns regarding privacy, as company-related projects cannot be openly disclosed.
   To address the issue mentioned above, I propose the idea of dividing LinkedIn users into two distinct age groups: users under 20 and users over 20. For users under 20, it is more likely that they have not yet worked with any companies. Therefore, I can assign a risk factor of 2 to the projects section of their profiles. However, for users over 20, there is a higher possibility of having company-related projects. In such cases, I recommend allocating a risk level of 4 to the projects section of their profiles to account for this increased risk.

10 Members Contribution

Aydin: Determining and analyzing the features of a legitimate company, qualitative analysis of the features, and extracting the top 5 important factors based on the survey. Generating the risk level for all data points by employing a reward function over the features in the dataset. Employing the multi-classification models to predict the risk level on the dataset and prepare the quantitative results from the model. Applying the SMOTE to the model to tackle the model imbalance challenge and utilizing the k-fold cross-validation to evaluate the

model's robustness.

Hrudaya: Conducted research on risk management approaches and best practices and gathered relevant information on various risk factors and their potential influence on the candidate's resume during the job application process. Utilizing the research, formulated a systematic framework or methodology to categorize and prioritize risk factors based on the impact. Performed risk assessments and examined historical data to enhance the analysis.

Kiran: I worked on dataset creation. I was primarily involved in data preprocessing and analysis. I performed tasks such as importing the dataset, handling null values, and checking for duplicates. I conducted exploratory data analysis (EDA) by creating visualizations to gain insights. Additionally, I implemented feature selection techniques using correlation analysis. I also played a role in training and evaluating the Decision Trees algorithm for classification. With the help of a decision tree algorithm, I predicted whether the company was real or fake.

An Tran: I was into the development and implementation of a scraper written in Python. The scraper generates a new version of the resume with masked information tailored to the specific company's risk factor criteria. Additionally, we discussed the trade-offs involved in using the scraper, both for the applicant and the recruiter, and mentioned areas for future enhancement, such as expanding the capabilities of the scraper to handle different resume formats and improving keyword detection.

References

1. Adikari, S., Dutta, K.: Identifying fake profiles in linkedin (2020), unpublished work
2. Advice, I.C.: Should you put your address on your resume? (2023), https://www.indeed.com/career-advice/resumes-cover-letters/should-you-put-your-address-on-your-resume, [cited 2023 May 21]
3. Blog, J.: Resume sections: How to organize and format your resume (2023), https://www.jobscan.co/blog/resume-sections/, [cited 2023 May 21]
4. Chron.com, W.: Importance of personal information on a resume (2023), https://work.chron.com/importance-personal-information-resume-20632.html, [cited 2023 May 21]
5. Corporation, S.: Privacy policy (2023), https://www.symplicity.com/compliance/privacy/privacy-policy, [cited 2023 May 21]
6. Documentation, S.: Spacy 101: Everything you need to know (2023), https://spacy.io/usage/spacy-101, [cited 2023 May 21]
7. Joshi, S., Nagariya, H., Dhanotiya, N., Jain, S.: Identifying fake profile in online social network: An overview and survey. In: Proceedings of the International Conference on Machine Learning (2020)
8. Khaled, S., El-Tazi, N., Mokhtar, H.: Detecting fake accounts on social media. In: Proceedings of the 2018 IEEE International Conference on Big Data (Big Data). pp. 3672–3681 (2018). https://doi.org/10.1109/BigData.2018.8621913

9. Pandya, J., PandiJain, G.: Detection of deceptive accounts using machine learning algorithm and deep neural network (2020), unpublished work
10. Python-docx Documentation: Python-docx Documentation. Latest Version (2023), `https://python-docx.readthedocs.io/en/latest/`, [cited 2023 May 21]
11. Sahoo, P., Lavanya, K.: Identification of malicious accounts in facebook. International Journal of Engineering and Advanced Technology (2019)