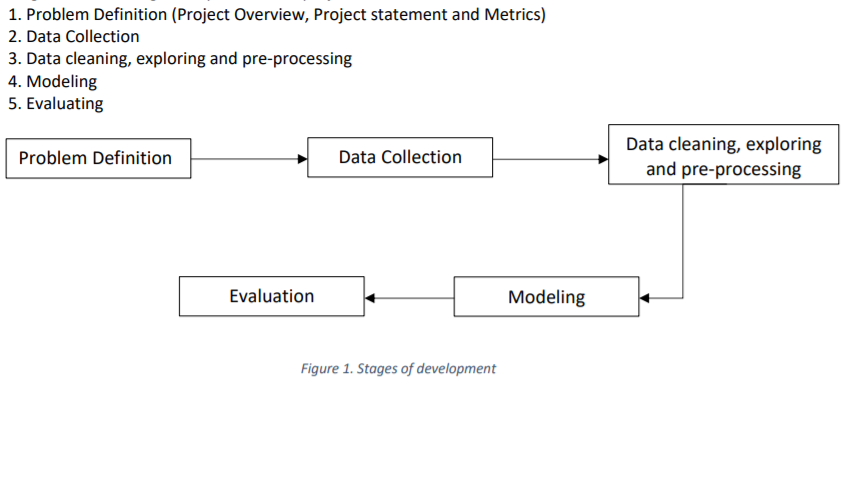
# 1. Introduction

There's been a marked increase in the number of scams related to employment. A 2018 CNBC study found that employment scams had doubled from 2017 levels. In the current economic climate, high unemployment is rampant. Unemployment levels are extremely high because of the epidemic, financial problems, and joblessness that the economy and its losses have brought to many people. This type of situation gives scammers a prime opportunity. Many people are getting taken advantage of because of this unprecedented incident, which has caused some amount of desperation. Most scammers will do this to steal someone's personal information. Information like home address, bank details, and social security numbers are all part of personal information. I'm in college, and I've gotten scammed many times by this. Scammers promise huge earning potential in exchange for monetary compensation. It is possible for a job seeker to be required to make an investment with the understanding that they will get a job as a result.

More scams have recently been detected in the job placement industry. Because of the market situation, significant numbers of people are unemployed. Many people have lost their jobs or have no job prospects due to the recession and the virus's effect. An opportune moment for committing fraud has arisen. When disaster strikes, it is only a matter of time before these swindlers target victims. Criminals typically use personal data from their victims to steal money or obtain other benefits. Data such as social security numbers, credit card numbers, and bank account information may be present. Job seekers have to launder money and pretend to work for the fraudsters for the opportunity to get a job. Job seekers could be forced to invest in fraudsters who will offer employment as a reward. Machines can analyze and process these kind of situations using algorithms.

Initially, I'll do data visualization to understand the results of the data, then use NLP to convert the text and use classification methods to divide job postings into two categories: fraud and the rest. We will look at a range of different measures to judge how successful our classifier is.

the five steps included in this project are as follows:



## 1.1.Data

The Kaggle website hosts the dataset used in this project. There are a complete of 17,880 sample points in this dataset, and it is split into two classes: the real data and the fake data. There is a bias because of the abundance of real job ads. There are only 800 fraudulent job postings while there are 17,880 genuine job advertisements. This data includes a grand total of 17 variables, all of which are independent, and it encompasses one parameter that is used to classify the entire data set. The seventeen factors are things like salary, title, location, etc.

We need to manually label the data that is going to be classified using AI so that we can train those same classifiers to identify fake job postings.

Combination data types make up the information. The following is a brief explanation of the variables:

Job ID: This characteristic explains about the numbers provided to each job to help in identification.

Title : the title says what the position is called

Location: The information in this feature shows the locations of the job.

Department :This feature provides information about the department.

Salary Expectation: Range of pay

company profile: It describes the company's background and history.

Description: an overview of the position. This feature describes about the position in the job.

requirements: To be eligible for the position

benefits : Work perks

telecommuting : Do office or remote work opportunities exist?

has\_company\_logo: is there a company logo on the job posting

has\_questions: Is the job posting asking any questions?

employment type: There are five different kinds of roles defined by this feature, including full-time, part-time, contract, temporary, and "other" workers.

required experience: Required experience can be (a student intern or entry-level hire), (a director or C-suite executive), or N/A.

required education: High school diploma, various associate degrees, high school education, unspecified college courses, any credential, vocational or high school coursework, or professional work experience are all acceptable degrees.

Industry : The industry for which the job entry is applicable

Function : The process of deciding a job's purpose

Fraudulent : Real or Fake: The target feature

## 1.2. Project Aim :

The goal of this project is to develop a classifier that can distinguish between real and fake jobs. The outcome will be assessed using two alternative models. Because the data contains both numerical and text elements, one model has been applied to the textual information and the other to the numeric data.

The end result will be a hybrid of the two. The final result will accept any appropriate job application data and generate a result indicating whether or not the position is authentic.

# 1.3. Project Objective:

* Use natural language processing (nlp) for lowercasing, stop words, stemming, and others to properly clean and pre-process text data.
* To analyze the text data, use NLP algorithms like Average Word2Vec and TF-IDF.
* Use several classifiers to distinguish between fraudulent and legitimate job advertisements.
* In order to get the greatest Accuracy from the classifiers, avoid overfitting and underfitting.
* Select the best model that provides the highest level of accuracy.

## 1.4. Tools

There are a number of tools employed as phase of this project to achieve its main objective. a few essential tools utilized during this Endeavour include:

Numpy : NumPy is a Python library that supports multi-dimensional arrays and provides a comprehensive collection of high-level mathematical operations for manipulating them. It's also commonly known as "numerical Python."

Pandas : Pandas is a tool for manipulating data analysis and data exploration in Python, and it's very useful for that.

Scikit-learn: Python's machine learning toolkit, offers a wide array of algorithms for the various tasks of classification, regression, and clustering.

Matplotlib: A Python-based data visualization and plotting package, is used to visualize and plot data.

Seaborn : The Seaborn data visualization library shares similarities with Matplotlib and it also provides some advance visualizations like heatmap for co-relation visualization.

NLTK: NLTK, which is short for Nlp Tool Kit, is widely used to pre-process text data. We can perform stemming, lemmatization, TF-IDF Vectorization by using Nltk.

# 1.5. Research Questions

A research topic is a topic that a study tries to find answers for. This refers to something in the study that's addressed by examining the data and telling the storey it reveals. The research objective is typically formulated in a way that highlights a variety of aspects, such as the study's research population and factors, and also the study's main purpose. Research is commonly centred around scientific research. It's not surprising that researchers frequently revisit and revamp their research questions: Research questions tend to be evolving rather than static. Researchers must reassess and adjust questions as they conduct literatures and build a framework for the study.

This initiative provided a number of new answers, including these to the following research questions:

Which of these two, TF-IDF or Word2Vec, do you think will give a superior performance?

How can we find the best machine learning algorithm for differentiating fraudulent and real job listings?

In what ways does balancing datasets affect classifier performance, and how significant is this for you?

## 1.6.Ethical, Legal, and Social Issues

Here we will discuss the ethical, legal, and social issues that could occur as the outcome of the project. We will discuss each of the identified threats that could occur in detail.

# 1.6.1. Ethical issues

Hateful & Criminal Actors:

The failure of accurate predictions of the model could result in passing the fake jobs that could reach the desperate job seekers and they could be cheated by the fraudsters.

The model that is being built should be highly accurate in predicting the fake jobs and should be robust to any outliers. It should work well on the unseen data.

Algorithmic Biases:

Algorithmic biases that might arise due to the imbalance in the data could result in poor performance of the classifiers on the real-world data. Such scenarios need to be avoided by the programmers while handling the data. The data should be balanced and then trained using the classifiers that could result in better performance of the algorithms.

# 1.6.2. Legal Issues

Failure to identify the fake jobs as claimed could get the owners in legal trouble. Hence it is highly essential to test the model on different scenarios, different types of new data and evaluate the model, make improvements in the model to make it more robust, and then it is recommended to deploy the model.

# 1.6.3. Social Issues

The corruption in the algorithm might misjudge the fake jobs as real and this could lure many job seekers and when the job seekers identify the reality there are high chances that they might be mentally depressed and face psychological issues. This could be a reason to harm individuals.

# 2. Methodology

## 2.1. Installing set-up

I used Python 3.7, which I downloaded from the official Python website, for my project. I installed Python on my machine by setting it up on my hard drive and adding Python to my path. Additionally, I've installed Anaconda on my computer to get everything ready for running Python via Jupyter. After that, I was able to execute this project by using command prompt to install a few libraries. Installing the libraries required the following commands:

pip install pandas

pip install numpy

pip install scipy

pip install scikit learn

pip install matplotlib

pip install seaborn

pip install nltk

pip install collections

pip install math

pip install collection

pip install warnings

## 2.2 Data Exploration

The data for this project is available at Kaggle - <https://www.kaggle.com/shivamb/real-or-fake-fake-jobpostingprediction>. The dataset includes 17,880 records with 18 variables combination data kinds make up the information. The following is a basic explanation of the variables:

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | Variable | Datatype | Description |
| 1. | job\_id | Int | Identification number given to each job posting |
| 2. | title | text | A name that describes the position or job |
| 3. | location | text | Information about where the job is located |
| 4. | department | text | Information about the department this job is offered by |
| 5. | Salary\_range | text | Expected Salary Range |
| 6. | Company\_profile | text | Information about the company |
| 7. | description | text | A brief description about the position offered |
| 8. | requirements | text | Pre-requisites to qualify for the job |
| 9. | benefits | text | Benefits provided by the job |
| 10. | telecommuting | boolean | Is work from home or remote work allowed |
| 11. | has\_company\_logo | boolean | Does the job posting have a company logo |
| 12. | has\_questions | boolean | Does the job posting have any questions |
| 13. | employment\_type | text | 5 categories – Full-time, part-time, contract, temporary and other |
| 14. | Required\_experiance | text | Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive or Not Applicable |
| 15. | Required\_education | text | Can be – Bachelor’s degree, high school degree, unspecified, associate degree, master’s degree, certification, some college coursework, professional, some high school coursework, vocational |
| 16. | Industry | text | The industry the job posting is relevant to |
| 17. | Function | text | The umbrella term to determining a job’s functionality |
| 18. | Fraudulent | boolean | The target variable à 0: Real, 1: Fake |

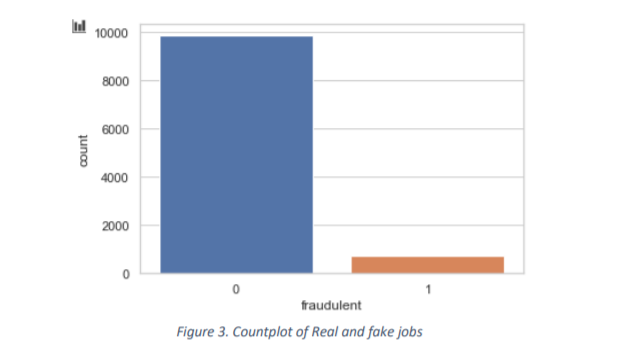
Here most of the data types are either Booleans(bool) or text a summary statistic is not needed here. The integer job\_id which is not appropriate for this analysis. The dataset is further explored to identify Exploratory data Analysis and Null values .



Figure2: Missing Values

In the above figure We can see that there are lot of missing values(missing information) in our dataset. We will handle these missing values in preprocessing stage.

The dataset is highly imbalanced with 9868 (93% of the jobs) being real and only 725 or 7% of the jobs being fraudulent. We can see the imbalanced dataset by using countplot. A countplot of the same can show very clearly.

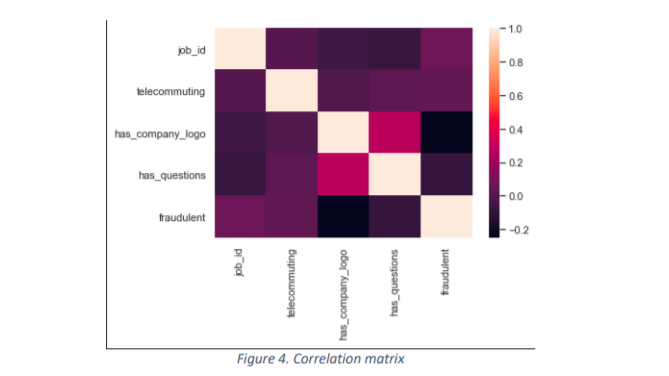


## 2.3 Exploratory Visualization

The main step of Data Science project pipeline is Exploratory Data Analysis. Because with the help of EDA we can visualize the data and find some insights of the data. Using summarize statistics and data visualizations, data analysis is a critical approach of the initial exploration of data to find gaps, contradictions, and assumptions. John Tukey's pioneering work formed the basis of the computational data analysis method Exploratory Data Analysis (EDA). Applied researchers use EDA because it enables a number of data analytic operations to take place and because it can support a multitude of different sorts of data and architectures. EDA incorporates graphic representations and interactive data visualization to perform model construction, analysis, and assessment of important distribution measurements.

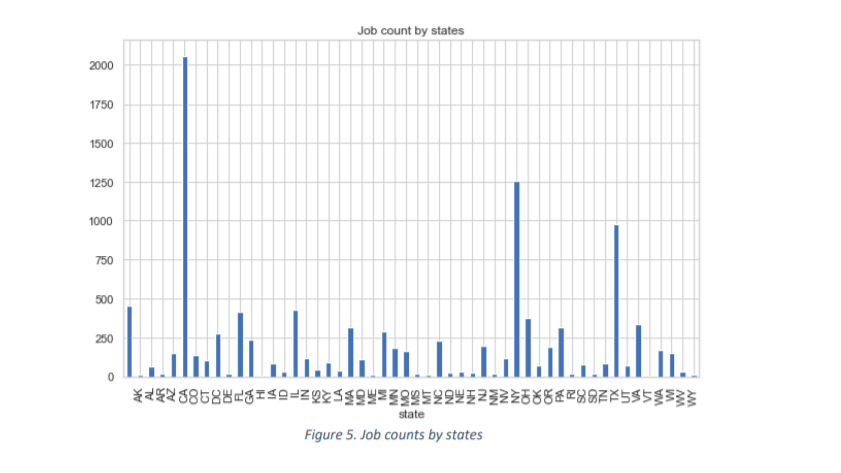
Data-based learning is at the heart of every study, and researchers use EDA's tools to refine the ways hypotheses are tested, so that important elements of data are discovered while modelling and estimating are being conducted. The initial stages of science are when hypothesis and model building is so important, and that's where the EDA (the effectiveness design award) comes in.

In order to investigate the relationships in the data, the first thing we must do is develop a correlation matrix.



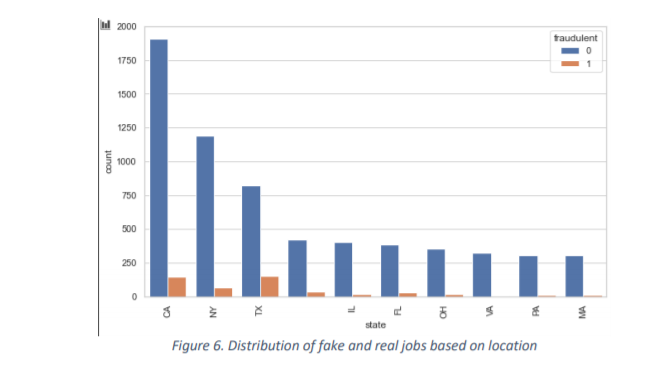
The correlation matrix illustrates that the numeric data show little to no correlation with one another. With telecommuting, a certain trend was found in relation to the Boolean variable. 92% of the time, a fraudulent job has a zero for this variable's value. To start, this dataset was analyzed for its textual content.

For this study, we begin with a consideration of place.



The graph at the top indicates which states create the most jobs. The three states with the most jobs posted are United states, New York, and Texas. Let's make another bar chart to better understand this.

In this barplot, you can see the accurate and inaccurate jobs inside the top 10 states' distributions.

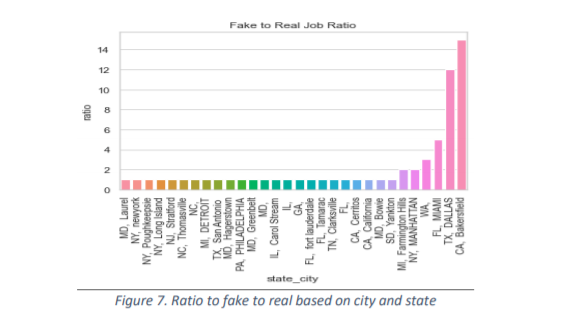


Comparing Texas and California to other states, there is a higher probability of fraudulent jobs in the former two. A new equation is created by including states as well as a ratio. A state and city-specific fake-to-real job ratio has been developed. In order to find out how many fake jobs exist for every real one, we use the following formula:

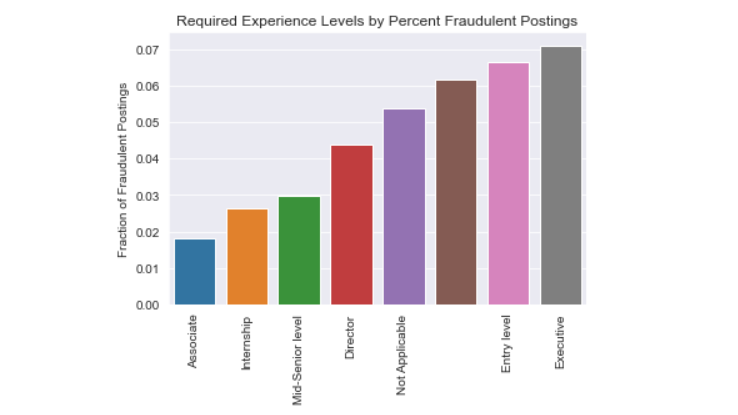
Ratio = state & city | fraudulent=0

State & city | fraudulent=1

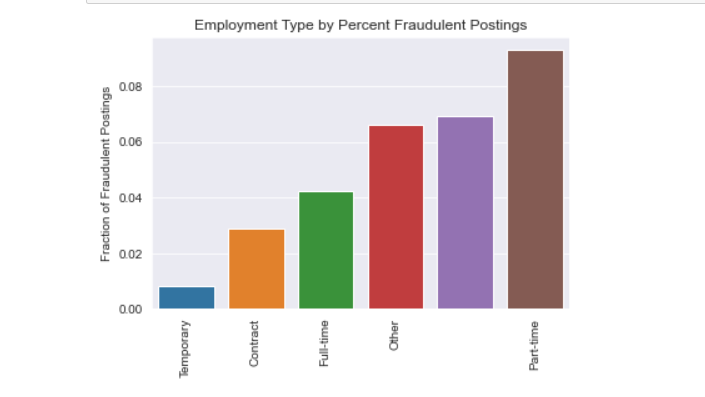
Only ratio values greater than or equal to one are plotted below.

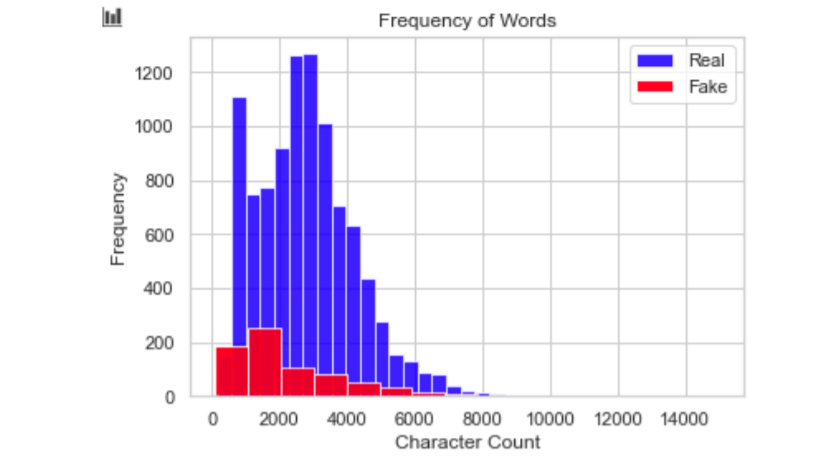


1) Job count based on Required experience



2) Job count based on Employment type

The text category is a field that integrates the various text-based fields under one umbrella for the purpose of broadening the discussion on the relationships between textual categories. Title, location, company profile, description, prerequisites, benefits, experience, education, industry, and function are the categories that this title belongs to. Histograms showing character counts for legitimate(Real) and malicious(Fake) jobs are visualized to show the differences. In looking at the frequency, it is clear that the fake jobs have a significantly lower character count.



## 2.4. Feature Engineering

An Attribute or a feature is a quality of a data that helps the user identify or predict the way it will act in a specific situation. As much as it contributes to the model's goal, any feature may be considered a feature. The role of a feature is made considerably obvious when you understand the feature's issue. Features are aspects of a product that can help overcome a business issue.

In order to help machine learning algorithms, you can use domain knowledge to develop features that assist in the process. Feature engineering in machine learning algorithms uses the raw data to create features that help the machine learning. The key component of the machine learning pipeline is feature engineering, which has a large effect on the machine learning model's efficiency.

The process of feature engineering has several approaches to create new useful features. Those are as follows:

1. Feature Selection
2. Feature Transformation
3. Feature Construction
4. Feature Extraction

**Feature selection:**

The process of feature selection entails selecting a subset of features from a vast array. Through focusing on the most critical functionality and condensing the feature set, computing of machine learning and data analytic algorithms becomes more feasible. Additionally, feature selection increases the efficiency of the performance generated by algorithms.

**Feature transformation:**

Feature transformation is the process of generating new features from existing data by the application of mathematical operations.

**Feature construction:**

Feature construction is the process of creating new variables in addition to those created during feature transformation that are relevant variables for the process under study.

**Feature extraction:**

The method of feature extraction is used to reduce the dimensionality of a dataset. The process of feature extraction entails merging current features to create new ones, thus decreasing the number of features in the dataset. This condenses the data into usable chunks for algorithms to handle without distorting the original associations or pertinent details.

On our data to build new useful features from the raw features, I have used the feature engineering technique. The approach used on our data is using the features name of the item and the item description, I have applied feature engineering method to extract the length of these features and created two new features that could help in improving the performance of the model. To be specific, the number of words in the text is counted and listed as new features for both name and item description.

For example, the name of an item listed on the platform is “Leather Horse Statues,” the length of the name would be “3.”

Similarly, the item description of an item listed on the platform is “The product is A grade with great performance,” the length of the description would be listed as “8.”

## 3.5. Data Cleaning

Data cleaning is critically important step in any Data Science/ Machine learning project. Cleaning up Data is crucial because dirty data leads to inaccurate results. It has a significant impact on the construction of a model. The most complex feature of machine learning isn't nearly as difficult as you might think, and there aren't any additional intricacies to learn about. Nonetheless, cleaning data correctly is what leads to or deters project success. Data scientists believe that investing a lot of time in the data step results in a better final algorithm.

When the dataset is thoroughly scrubbed, we can get pretty good results using simple algorithms, which is especially useful when the dataset is large and computation is a concern.

**Steps involved in Data Cleaning:**

3.5.1.Handling missing data

Missing data can be hard to spot if you're using machine learning. Ignoring or erasing the lack of information isn't a viable option. It is imperative that they be handled delicately, as they can indicate an important matter. Two ways to handle missing data are widely used:

1. Dropping the rows containing missing values
2. It is not advisable to drop missing values because doing so reduces information.
   * Perhaps the missing value is telling us something.
   * In addition, you often have to make predictions about new data even though some of the attributes are missing in the real world!

3 ) Deriving a series of past observations to assist in inserting missing values.

Trying to assign missing values is a bad idea because no matter how advanced your imputation technique is, filling something in that was originally missing will lead to information loss.

* + The absence of a value almost always has a meaning, and your algorithm should know if a value is missing.
  + You haven't added anything of value by building a model to assign your values. You're just perpetuating the existing features' patterns.

## 3.5.2. Encoding of Categorical Features

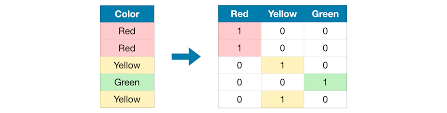
A category variable does have two or more categories. A categorical variable can be either nominal or ordinal. Categories of a nominal variable do not have any built-in ranking. Gender is often a categorical data that is simply composed of two categories with no natural order. The ordinal variable exhibits a well-defined order.

Many machine learning algorithms can't understand the categorical data on their own. On the other hand, Decision trees can learn straight from the data itself. They require all parameters to be numeric because of this. Categorical data needs to be transformed into numerical data.

Few types of categorical variable encoding are:

**1 ) One hot encoding**: In one hot encoding the categorical variables will be converted into the 0 and 1 form . 1 is use if the category is present and if the category is absent then 0.

Example : We can see in the figure In the first row red is present, so In the first row only red column is 1. Others will be 0.



One-Hot Encoding

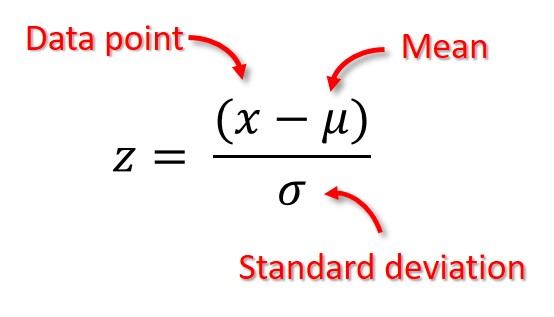
2. **Integer Encoding / Label Encoding**: In Label Encoding simply we will give the numbers to each category .In this project I have done Label encoding for converting the categorical features into numerical features.

## 3.5.3 Feature Scaling

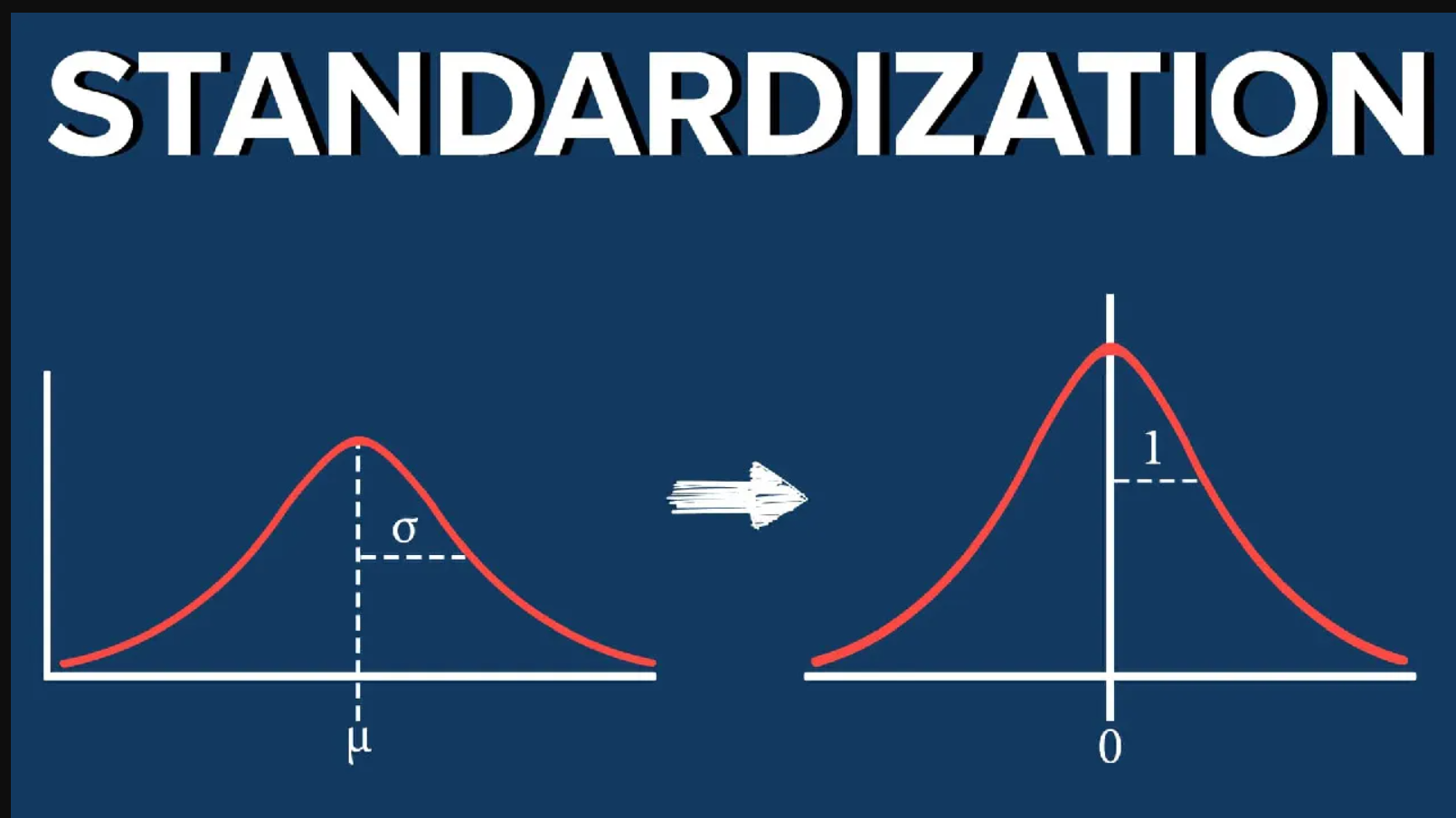
Data normalization is the process of converting information to a standard format to enable the user to process and analyze it. Data standardization helps in the improvement of your data by trying to transform and standardizing it.

Data Standardization means we want to convert the data in the particular range (0 to 1) . so that it is easy for the algorithm to handle this data. Data Standardization is very helpful in case of distance based Algorithms( KNN, K-Means).

We can standardize the data with the help of z-score method.



z- score scales the data range between between 0 and 1.



## 3.6. Pre-processing

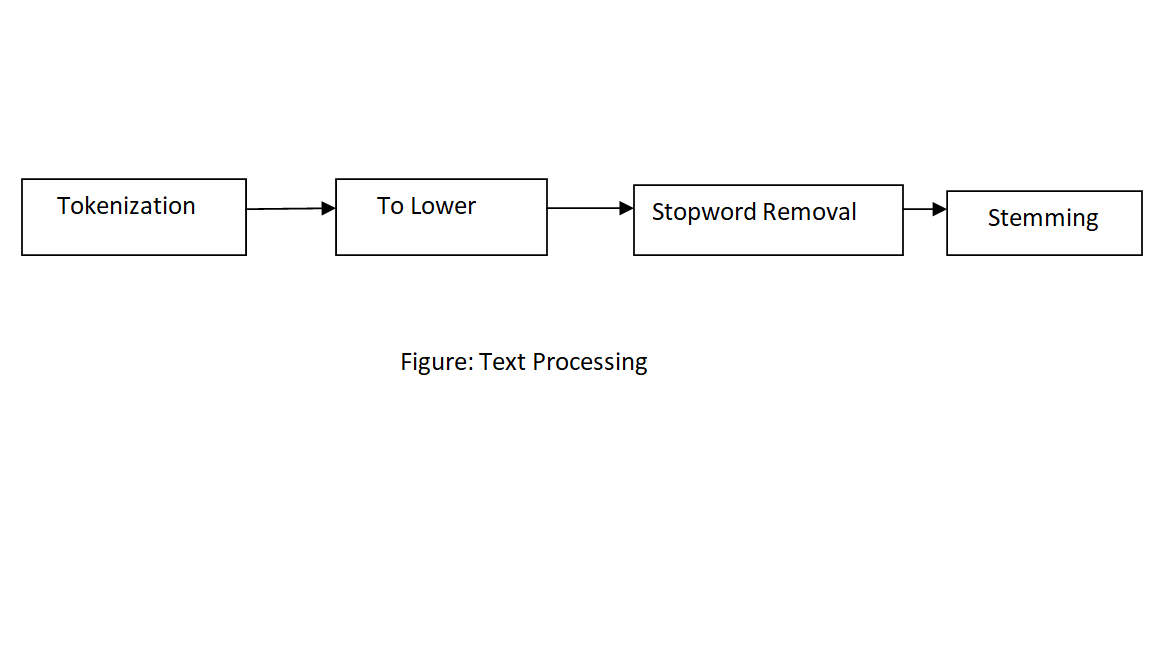
Pre-processing data is a critical stage in the data science pipeline. It is a data science technique that converts unstructured data to a usable format. Raw data (data from the real world) is unstructured and cannot be processed directly. Therefore data must be pre-processed prior further process in data science pipeline.

The initial pre-processing techniques that were used as part of this project is to import all the libraries that were installed and load the dataset using the Pandas library. Further, the size of the data, the variables in the data, understanding the types of variables were performed as part of the pre-processing. To add further, the missing values were checked such that to eliminate the missing values in the data. The missing values in the data adds no weight to the data.

## 3.6. 1. Text processing for Machine Learning

Text processing is one of the important tasks when we have text variables in our data. The text cannot be feed to a machine learning algorithm directly. For the text to be processed by the machine learning or deep learning algorithms text processing should be done on the text variables.

Text processing is a natural language processing technique that is used to transform the text data into numerical variables so that they can be processed by the algorithms. There are many steps that are involved in text processing.



3.6.2. Text pre-processing

Text pre-processing is a stage where certain operations are performed on the text to process it to further steps. Pre-processing the text literally involves transforming it into a predictable and analyzable format for that is suitable for our task. Few of the operations include lower casing the text, removing punctuations, etc., there are explained below.

**Lower-casing:**

Lower casing is a standard pre-processing step that is performed on the text data. The function of the lower casing is to convert the text in the data to lower cases.

Example:

The sentence “What are you doing?”

Will be transformed as “what are you doing?”

**Removing Punctuations:**

Removing punctuations is another step in pre-processing technique that handles the punctuations in the text data. Punctuations are the special characters such as ‘;’‘,’, ’.’, ’!’, ’$’, ’#’, etc., that are included in the data.

Example:

The sentence “What are you doing?”

Will be transformed as “what are you doing””

**Removing Stop-words:**

In the English language there are certain common words that do not have semantic are considered as stop words. Words such as ‘a’, ‘the’, ‘them’, ‘they’, ‘I’, ‘is’, ‘as’, etc., are considered as stop words.

Example:

The sentence “these are the balls in the basket”

Will be transformed as “ball basket”

**Stemming:**

Stemming is a process of transforming the natural English language words to their root words by dropping suffix from the words.

Example:

The word “studying”

Will be transformed as “study”

## 3.6.3. Handling text features and categorical text features.

After the pre-processing has been performed on the text data the further step is to feature the text features using natural language processing techniques such as Term Frequency – Inverse document frequency for the text features and count vectorizer for handling the text categorical features.

### 3.6.3.1. TF-IDF for text features

TF-IDF, also known as Term frequency – Inverse document frequency can be defined as a natural language processing technique to convert the words to numerical vectors that are less sparse by computing the relevance/importance of a word in a document.

Term frequency - The logic is that if a term appears several times in a text, its value can be increased since it is more important than other terms that appear less often (TF).

Inverse Document Frequency - If a term appears more often in a text document and also in a large number of other documents, it is possible that this is because the term has a regular occurrence; not that it is important.

In our project, the term frequency and inverse document frequency is used to handle the text features such as description.

## 3.6.3.2. Avg Word2 Vec for text features

Average Word Vectors can be used to create embeddings for documents, paragraphs, and sentences. It is a common technique for many NLP applications to use word vectors in projects that are aiming to do larger scale text projects, such as in documents, paragraphs, or sentences.

## 3.7 Balancing Dataset

As we know the given dataset is imbalance dataset. So, we have to balance it before giving it to ML Model. I have used Smote for balancing the dataset.

#### SMOTE: Synthetic Minority Oversampling Technique

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class.

This algorithm helps to overcome the overfitting problem posed by random oversampling.

## 3.8. Algorithms and Techniques:

For this project I have used Machine Learning algorithms with TF-IDF and Avg Word2Vec. These are the following algorithms which I have used in the project:

Logistic Regression :

The logistic regression method uses historical data to forecast the value of a new data point. In logistic regression, dependent data is predicted by analyzing how a single independent variable or multiple independent variables are related to one another.

Logistic regression is applied to find the odds ratio when multiple explanatory variables are present. The only difference is that the dependent variables is binomial instead of the usual multinomial. The outcome is the implications of each parameter on the probability ratio of the measured interest event

# Naïve Bayes Classifier:

The Naive Bayes is a supervised tool that uses the Conditional Probability concept in Bayes Theorem. Even though this classifier is inaccurate, its decision is extremely effective. The classifier achieves an exceptional result when the attributes are independent, or when the features are completely dependent on their functionality.

This classifier is only dependent on information loss.

# K-nearest Neighbour Classifier

K-NN Classifiers, recognised as lazy learners, identify objects in the feature space by selecting examples with the most similar feature set. The classifier examines a total of k objects and picks the closest one to assign to a class. Choosing the value of k is a big problem with this classification technique.

# Support Vector Machine

SVM is a linear model that can be used for classification and regression. It can resolve both linear and non-linear problems, and it excels in addressing many common situations. The SVM concept is rather straightforward: A line or hyperplane that segregates the data in to the classes is generated by the algorithm.

# Decision Tree Classifier

Decision Trees (DTs) employ a tree-like structure to make classifications. It understands how to categorise. The test includes target classes that are denoted as leaf nodes of DT, and decision nodes that are non-leaf nodes of Decision tree that are used to indicate tests. Decisions will be able to be made using results that were either gathered by using one of those branches or that originated with that decision node. This tree starts from the base and finishes at a leaf node. It describes how classification results are attained by a decision tree.

Decisions trees have been successfully used in spam filtering. Using and training the whole model will allow for goal prediction through implementing a model to match some criteria.

# Ensemble Approach based Classifiers

The use of ensemble approaches improves machine learning algorithms by having them perform better in unison.

For classification problems, random forest (RF) uses ensemble learning methods and regression algorithms. This classifier has several tree-like classifiers in it, and these trees are given various subsamples of the dataset to identify the class of the input. Each tree casts a vote for the best class, and the most popular class gets selected.

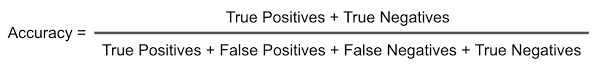
Boosting is a common technique in which learners with lower accuracies are integrated into a single higher-accuracy learner. To implement classification algorithm, the boost technique finds weighted majority votes using the reweighted training data and then classifies it.

Example: XG-boost

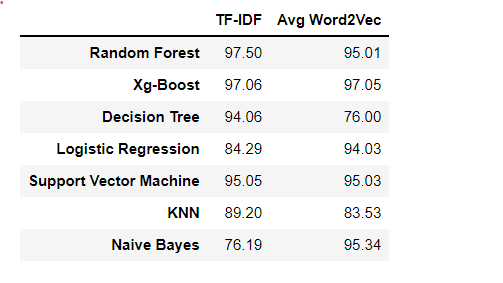
## 3.9. Performance Evaluation Metrics

While it is important to validate an evaluation of a model's skill, using metrics is necessary to evaluate the performance of a model. Metrics are employed to determine the most suitable problem-solving method. These metrics are what matter most.

Accurately identifying the ratio of correctly-predicted instances to the total amount of instances evaluated is a metric for measuring accuracy. While the accuracy might be adequate as a measurement for model performance, it may not be good enough because it fails to include incorrect predictions. If someone treats a fake post as a real one, it could cause a serious issue. False positives and false negative issues that accommodate for misclassification should be considered because of this.



Here I have compared the Accuracy of all the algorithms with TF-IDF and Avg Word2Vec form. We can see that Random Forest Algorithms gives higher accuracy with TF-IDF.



# Conclusion

Jobseekers will be advised about where to look for legitimate employment by being warned about jobs scams. Multiple machine learning algorithms have been proposed as countermeasures for employment scam detection in this paper. In the context of scam detection, supervised mechanism has been employed to highlight the implementation of a number of classifiers. The research has shown that Random Forest algorithm outshines its competitors.

In this paper, I have worked with different models of Machine Learning to classify the fake job postings from the given data to identify the better performing models.

Initially, the exploratory data analysis was performed to understand the insights in the data. In Exploratory data Analysis I have performed various data visualization in order to understand the data. Further, feature engineering was performed to extract new features based on two of the features in the data. Additionally, the text processing is performed as per the requirements of the models. I have then trained 7 Machine Leaning models .To evaluate these models, I have used accuracy as the metric. By analyzing the performances of the models, it has been observed that the Random Forest with TF-IDF has recorded the Highest accuracy. The performances of the models are quite similar and do not have much difference irrespective of TF-IDF and Avg Word2Vec.

To answer the research questions posed, after performing the analysis on the data with different models with all the evidence we can conclude that the TF-IDF and Avg Word2Vec models are performing similarly with a slight difference in the performances. There is no drastic difference in the performance of the models except Naïve Bayes. Because Naïve Bayes gives 76.19% accuracy with TF-IDF and 95.13% accuracy with Avg word2vec. It has been observed that Xg-Boost and Random Forest gives highest accuracy. Artificial examples are generated as opposed to copying the instances and overfitting, a common problem when using random oversampling, is prevented by balancing datasets. There are no problems with important data getting lost.

Reference Images:

<https://medium.com/@joelvarmadirisam/my-take-on-why-do-we-need-to-standardize-the-data-1459e6608a63>

<https://www.geeksforgeeks.org/>