Natural Language Processing- Detecting Real or Fake Job Posting

University of Derby

100568146@unimail.derby.ac.uk

Abstract.
This paper conducts a detailed data analysis report on detecting real or fake job postings on internet. The properties of Natural Language Processing is done to determine whether the job posting is real or fake. Several models are tried and the best model to classify the classes are determined using the performance of these models. A deep learning algorithm using CNN and RNN is also trained with respect to the job description and the labels.

Keywords: Natural Languge Processing, Random Forest, Naïve Bayes, Support Vector Machine, Convolutional Nueral Network, Reccurent Neural Network, LSTM , Text-summary.

Contents

1 Introduction

Natural Language Processing is a subset of Artificial Intelligence where the human language is made understandable by a system in other words it is way to make human language intelligible to machines [Monkeylearn,2021]. This technology uses several applications to breakdown the speech data to make it meaningful and thus to process accordingly [GeeksforGeeks, 2021]. This paper is a detailed report of NLP in detecting fake and real dataset.

Aim and objectives.
As a graduate or for anyone else looking for a job it would be helpful if one can get an idea if a given job posting is fake or real. The goal of this NLP project is to develop a training mechanism that can identify a real or fake job posting in internet.
The objective of the study is a model to predict real and fake job postings.

2 Dataset

The data set used in the study is the real/ fake job postings from Kaggle [Kaggle]. The data set contain 18,000 job descriptions and have 18 variables. Most of the variables in the dataset is textual information.

2.1 Data Variable Description.

Figure 1 shows a small summary of all the variables in the dataset and the variable type. The dependent variable or labels is the column fraudulent which takes two values 0 and 1. 0 represents a real job and 1 represent a fake job posting.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   job_id               17880 non-null  int64
 1   title                17880 non-null  object
 2   location             17534 non-null  object
 3   department           6333 non-null   object
 4   salary_range         2868 non-null   object
 5   company_profile      14572 non-null  object
 6   description          17879 non-null  object
 7   requirements         15185 non-null  object
 8   benefits             10670 non-null  object
 9   telecommuting        17880 non-null  int64
 10  has_company_logo     17880 non-null  int64
 11  has_questions        17880 non-null  int64
 12  employment_type      14409 non-null  object
 13  required_experience  10830 non-null  object
 14  required_education   9775 non-null   object
 15  industry             12977 non-null  object
 16  function             11425 non-null  object
 17  fraudulent           17880 non-null  int64
dtypes: int64(5), object(13)
memory usage: 2.5+ MB
```

Fig. 1.

3 Data Pre-processing

The dataset contains 18,000 job descriptions out of which 866 are fake. Figure 2 shows the distribution of fake and real jobs from the data set. Clearly the data set is not equally distributed. The data pre-processing starts with dealing the imbalanced data.

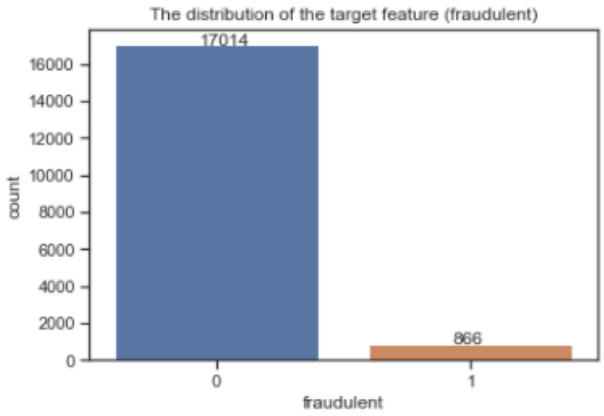

Fig. 2.

3.1 Balancing the data

It was observed that the data is highly imbalanced and for the further study the data considered is a subset of the original data where equal number of fake and real job postings where considered. The new data has a distribution as follows in the figure 3.

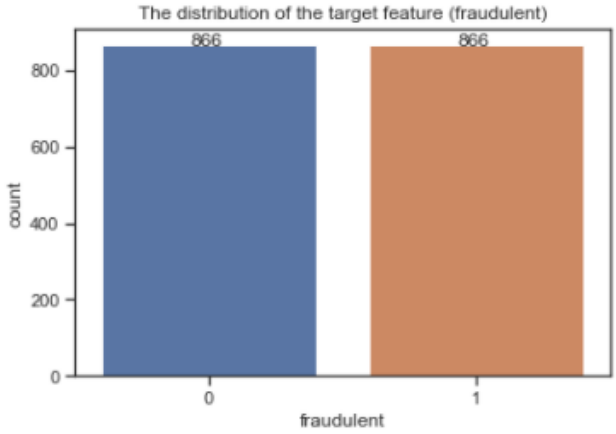

Fig. 3.

3.2 Removing all the missing values

The variable of interests in this study are ‘description’, ‘requirements’ and ‘company profile’. The rows with a missing value in these columns were deleted using the ‘dropna’ function. Hence the new data is a subset not containing any missing values in these columns.

3.3 Data Pre-processing for NLP

Prepping the dataset for NLP is an important step before moving on to build any models. The variable of interest here is ‘description’ in the first dataset and the second dataset is the ‘full’ variable which is a combined version of three of text variables from the original data namely- ‘company_profile’, ‘description’, and ‘requirements.

These texts were converted by removing all other variables from the texts and then converting every word to lower case letters and then removing the stopwords.

The resulted text data is used for data analytics part.

3.4 Sub-setting and dividing into train and test datasets

Diving the dataset into train and test is a necessary before each model building since it allows as to check the model performance. The data with the use of ‘train_test_split’ is divided into train and test in a ratio of 70 :30

4 Data Analysis

The data analysis part can be divided into two parts. The first part is to build the model using the ‘description’ to predict the job posting as fake or fraudulent and the second part is to combine the columns ‘description’, ‘company_profile’ and ’requirements’ and then to compare the model performances.

4.1 Classification algorithms in Natural Language Processing

Classification is a process of extracting information from a pre-existing training dataset to train the machine to identify different classes or categories for these training datapoints. The basic idea is to train a model so that it can assign any new datapoint into one of the classes [Wolff, 2020] that it has identified.

Classification algorithm is a kind of supervised learning method that is used to identify new observations based on already available data. The already available data is taken as the training dataset and the algorithm is trained using the labels for a particular category. In a classification algorithm the output is a particular category.

Naïve Bayes.
The naïve Bayes algorithm is a mostly used supervised learning algorithm that classifies the data points on conditional probability values. It is one of the important classification algorithms for text classification. It is a probabilistic classifier which predicts the probability of an object being in a class. The intuition of naïve Bayes is the Bayes theorem. [GeeksforGeeks, 2021].
The Bayes theorem can be stated as:

$$P(A/B) = \frac{\mathrm{P(B/A)*P(A)}}{P(B)} (1)$$

Equation 1 gives the probability of an event given some information. The naïve Bayes algorithm equivalently uses a certain prior information of a data point to classify its class.

Naïve Bayes algorithm for fake or real job posting data

A naïve Bayes algorithm to classify a job posting as real or fake is used. The technique of a naïve Bayes algorithm is to use the words in the given text to predict the class it belongs to. The idea of bag of words representation is to represent the text as a set of words and its counts- the number of times the word occurred in the text. Other information like the order of the words, the style of sentences is ignored.

The results of Naïve Bayes algorithm on both the datasets are given below: -

Naïve Bayes for	Accuracy
Description alone	0.76
Company profile, Description, Requirement combined	0.90

Table 1.

Table 1 clearly suggests there is a high accuracy increase in the combined model than using the description alone.

Random Forest.

Random Forest classifier is a supervised learning algorithm that can be used for regression and classification problems [Vadapalli, 2021]. The random forest uses multiple decision trees performed on bootstrap samples on a random basis and get predictions on each tree. The final class is decided by the majority votes from the different decision trees.

Random Forest for fake or real job posting data

A random forest classification on the datasets both original and combined dataset is performed to predict the job posting as fake or real. The result of random forest is combined in the following table. Table 2 suggest random forest as a very good classifier for classifying real and fake job postings with high accuracy.

Random Forest for	Accuracy
Description alone	0.85
Company profile, Description, Requirement combined	0.96

Table 2.

Random forest classifier also increased its performance after combining the columns.

Support Vector Machine

Support Vector machines are supervised learning algorithms which learns from the train dataset to classify a new point according to a certain mathematical criterion. The train dataset is mapped into classes by maximizing the width of the gap between the classes. And a new point is assigned to a class based on which side of the gap it belongs to [Wikipedia, 2021].

SVM for fake or real job posting data.

A support vector machine for classifying real and fake job posting for both the dataset gives the following results in table 3. SVM had a poor performance with accuracy 50% on description alone and it has a better performance of 67% on the combined dataset.

SVM for	Accuracy
Description alone	0.49
Company profile, Description, Requirement combined	0.67

Table 3.

5 Deep Learning

5.1 Word Embedding

Word embedding is the most important part in a natural language processing. The word embedding technique is a numerical representation of words such that a computer understands the text. It is a representation of words that can capture the meaning, context and find any relationship.

The most important step in any NLP project is the text preprocessing converting the text into some vector format so that the algorithm can understand these texts for any future predictions. Word embedding helps us o convert the text into a vector which can be understood by a system. A model is built for the variable 'description' using Word2Vec and each word is represented as a vector representation. Figure 4 represent the vector representation of the word 'team'.

```
Out[191]: array([-0.8233659 ,  0.8492673 , -0.9908197 , -0.615801  ,  0.10664734,
                 -0.7800181 , -0.01601642,  2.0880082 , -0.553629  ,  0.1682502 ,
                 -0.877015  , -1.1511708 ,  0.0180861 , -0.13483015, -0.7430255 ,
                 -1.0913174 ,  0.10242778, -1.0811269 , -0.18726072, -0.94161063,
                 -0.11195157, -1.5616312 ,  0.22131789, -0.14184982, -1.2046589 ,
                 -1.1584297 ,  0.9422437 , -0.63687694, -0.55436754,  0.6275539 ,
                  1.3219808 ,  0.16780776,  0.7492731 , -0.40702444, -0.3784161 ,
                 -0.6455123 ,  0.37985882, -1.505306  , -0.23442388, -1.4880016 ,
                  0.08529652, -1.2159128 ,  0.21603048, -0.06048988,  0.36238086,
                 -0.5911272 ,  0.79870325, -0.3563851 , -0.6884896 , -0.1304346 ,
                  0.9418084 , -0.7519797 ,  0.14045131,  0.29040083, -0.3356567 ,
                 -0.60762805,  1.2025362 ,  0.03270775, -1.1219065 ,  0.5802209 ,
                  0.3733196 , -0.71947247,  0.29553255, -1.100275  , -1.6312759 ,
                  0.10385704,  1.9977578 ,  0.8017517 , -1.167495  ,  0.99597335,
                  0.11535754, -0.4344377 ,  0.52934295,  0.5695438 ,  1.5498251 ,
                  0.8831911 , -0.13497941,  0.53662425, -1.1218303 ,  0.89774287,
                 -0.6638029 ,  0.41644982, -0.3629379 ,  1.2415437 , -0.09403069,
                 -0.4517005 ,  0.34019428,  0.15899117,  1.0882598 , -0.12975867,
                  1.4536324 , -0.5131831 ,  0.81057525, -0.31058112,  1.5094908 ,
                  1.2115775 ,  0.67948747, -1.6100545 ,  0.5253295 ,  0.3263358 ],
                dtype=float32)
```

Fig. 4.

The model can be used to find the most similar word. For example, figure 5 represents the most similar words of the word ‘work’.

```
Out[183]: [('orient', 0.9648072123527527),
           ('environ', 0.9625198245048523),
           ('abl', 0.9616612195968628),
           ('independ', 0.9595454335212708),
           ('ideal', 0.9549047946929932),
           ('must', 0.9515144228935242),
           ('taskabl', 0.9462124109268188),
           ('part', 0.9454842805862427),
           ('multitask', 0.9426288604736328),
           ('pace', 0.9417974948883057)]
```

Fig. 5.

5.2 Convolutional Neural Network (CNN).

Neural networks are algorithms inspired by neurons in animal brain. An artificial neural network consists of many nodes that takes several inputs and produce an output based on a certain threshold value. This threshold value is usually determined by an activation function. An artificial neural network is trained by adjusting the wait associated with each connection. A typical artificial neural network in a dep learning setting consists of many layers of connected nodes which progressively represents data in increasing complexity.

CNN in fake or real job posting data.

A convolutional neural network is built for predicting the fake and real job descriptions.

The first step is to use embedding technique to the text variable. Each variable is given a unique index in the complete vocabulary. The whole text is tokenized using the Tokenizer from the Tensorflow package. The tokenized test and train datasets are converted to sequences using the bag of word method. Each word is converted into an array of equal length using pad_sequences.

A five-layer CNN is built and used for prediction. For the given model ‘epoch’ is defined as five, since from number greater than five the validation accuracy starts decreasing that means the weights will be changed five times. After five iterations the model has a training accuracy of 98% and the validation accuracy is 87%. The model is a good fit for future references.

5.3 Recurrent Neural Network (RNN).

A recurrent Neural Network is built for the job description to predict real or fake classes.

A recurrent neural network is a class of artificial neural network which has memory that is it remembers all the information that has been calculated. RNN allows to operate over vector sequences. RNN has a short-term memory and there is a possibility that RNN can miss some information from the beginning.

Long Short-Term Memory helps in regulating the information for each neuron. The LSTM mechanism can decide which information to be considered as relevant and which not. The main idea of LSTM is that the output from any neuron is used as an input for the same neuron.

The model was built with iterations and a hyperparameter of thirty. The model is iterated over ten steps. The resulted model has a training accuracy of 0.55 and a validation accuracy of 0.54.

6 Text Summary

An automated text summary is created using the spacy library. The idea is to give each sentences a score after a sentence tokenization. The highest scored sentences will be taken as the summary of the texts.

The first step is to find the word frequency of each word in the texts and can be used to score each word. The score of each word is added up in a sentence and the highest scored sentence is taken as the summary.

For text summary the first entry from the description is taken as the text

```
Organised - Focused - Vibrant - Awesome!Do you have a passion for customer service? Slick typing skil
ls? Maybe Account Management? ...And think administration is cooler than a polar bear on a jetski? Th
en we need to hear you! We are the Cloud Video Production Service and opperating on a glodal level. Y
eah, it's pretty cool. Serious about delivering a world class product and excellent customer service.
Our rapidly expanding business is looking for a talented Project Manager to manage the successful del
ivery of video projects, manage client communications and drive the production process. Work with som
e of the coolest brands on the planet and learn from a global team that are representing NZ is a huge
way!We are entering the next growth stage of our business and growing quickly internationally.  There
fore, the position is bursting with opportunity for the right person entering the business at the rig
ht time. 90 Seconds, the worlds Cloud Video Production Service - http://90#URL_fbe6559afac620a3cd2c22
281f7b8d0eef56a73e3d9a311e2f1ca13d081dd630#90 Seconds is the worlds Cloud Video Production Service en
abling brands and agencies to get high quality online video content shot and produced anywhere in the
world. Fast, affordable, and all managed seamlessly in the cloud from purchase to publish. 90 Seconds
removes the hassle, cost, risk and speed issues of working with regular video production companies by
managing every aspect of video projects in a beautiful online experience.  With a growing network of
over 2,000 rated video professionals in over 50 countries and dedicated production success teams in 5
countries guaranteeing video project success 100%. It's as easy as commissioning a quick google adwor
ds campaign.90 Seconds has produced almost 4,000 videos in over 30 Countries for over 500 Global bran
ds including some of the worlds largest including Paypal, L'oreal, Sony and Barclays and has offices
in Auckland, London, Sydney, Tokyo &amp;amp; Singapore.Our Auckland office is based right in the heart of
the Wynyard Quarter Innovation Precinct - GridAKL!
```

Fig. 6.

Figure 6 gives us the text which we are going to summarize. The summary is created by pertaining 30% of the information from the original text. The thirty percent of the length of the original text is calculated which is five in this case and the summary is obtained in finding the five highest scored sentences. Figure 7 gives the summary of figure 6 in five sentences.

"\xa090 Seconds, the worlds Cloud Video Production Service -\xa0http://90#URL_fbe6559afac620a3cd2c222
81f7b8d0eef56a73e3d9a311e2f1ca13d081dd630#90 Seconds is the worlds Cloud Video Production Service ena
bling brands and agencies to get high quality online video content shot and produced anywhere in the
world.\xa0With a growing network of over 2,000 rated video professionals in over 50 countries and ded
icated production success teams in 5 countries guaranteeing video project success 100%.\xa090 Seconds
removes the hassle, cost, risk and speed issues of working with regular video production companies by
managing every aspect of video projects in a beautiful online experience.Our rapidly expanding busine
ss is looking for a talented Project Manager to manage the successful delivery of video projects, man
age client communications and drive the production process.It's as easy as commissioning a quick goog
le adwords campaign.90 Seconds has produced almost 4,000 videos in over 30 Countries for over 500 Glo
bal brands including some of the worlds largest including Paypal, L'oreal, Sony and Barclays and has
offices in Auckland, London, Sydney, Tokyo & Singapore."

Fig. 7.

7 Conclusion

A detailed Natural Language Processing is used to classify a job posting on internet as real or fake. Several models were built, and its performances were compared. The analysis was done simultaneously for two different datasets, one for the original description vs fraudulent and the second for a combined column of three of the variables in the dataset vs fraudulent.

Naïve bayes algorithm has an accuracy of 0.76 for the original dataset and an accuracy of 0.90 for the combined dataset. The NB algorithms accuracy is increased by almost 10% between the models built using the two datasets.

Random Forest classifier has an accuracy of 0.85 and 0.96 respectively for the two datasets. With these high accuracy rate among other models' random forest is the best fit for classifying a real and fake job description.

A SVM created on the two dataset has an accuracy of 0.49 and 0.67. SVM turned out to be the least accurate performing model in this case.

Two of the deep learning techniques, CNN and RNN were trained to classify real and fake job postings. CNN has a 0.87 accuracy whereas RNN has a low accuracy of 0.54.

The different model performances shows that there is a hope for future studies and that the built models can be used as a baseline for comparison and to use for predictions.

8 References

1. Vadapalli, Pavan., (2021), 'Random Forest Classifier: Overview, How Does it Work, Pros & Cons', Available at:- Random Forest Classifier: Overview, How Does it Work, Pros & Cons | upGrad blog (Accessed 9 January 2022)
2. Wolff. Racheal., (2020),' 5 Types of Classification Algorithms in Machine Learning', Available at: 5 Types of Classification Algorithms in Machine Learning (monkeylearn.com) (Accessed 24 December, 2021).
3. Wikipedia, 2021, 'Support-vector machine', Available at :- Support-vector machine - Wikipedia (Accessed on: - 1 January 2022)
4. GeeksforGeeks, (2021),'Natural Language Processing - Overview 'Available at:- Natural Language Processing - Overview - GeeksforGeeks Accessed on : 2 January 2022).
5. MonkeyLearn, (2021),' Natural Language Processing (NLP): What Is It & How Does it Work? , Available on:- Natural Language Processing (NLP): What Is It & How Does it Work? (monkeylearn.com) (Accessed 5 December 2021).
6. Silaparasetty, Vinita., (2020) 'Guide to Handling Missing Values in Data Science' Available at: Guide to Handling Missing Values in Data Science | by Vinita Silaparasetty | Medium (Accessed: 2 December 2021).
7. GeeksforGeeks, (2021), 'Naive Bayes Classifiers - GeeksforGeeks', Available at: Naive Bayes Classifiers - GeeksforGeeks (Accessed: January 2022)
8. Chakure, Afroz., (2019) 'Decision Tree Classification', Available at: Decision Tree Classification. A Decision Tree is a simple… | by Afroz Chakure | The Startup | Medium (Accessed :10 December 2021).
9. Kaggle, 'Real / Fake Job Posting Prediction Dataset of real and fake job postings', Available at :- Real / Fake Job Posting Prediction | Kaggle, Accessed on 9 January 2022.
10. S. U. Habiba, M. K. Islam and F. Tasnim, (2021) "A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Available at:- A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques | IEEE Conference Publication | IEEE Xplore, Accessed on 2 January 2022.