UNIVERSITY OF LEEDS

Machine Learning for Identifying High-Risk Online Job Advertisements

Preeti Sharma

Data Science Development Programme 2023-24

Leeds Institute for Data Analytics

March 2024

Millions of individuals worldwide are subjected to exploitative labour practices, highlighting the urgency of addressing modern slavery Organisation 2022. This project focuses on intervention at the recruitment stage, where deceptive job advertisements often proliferate unchecked. We adopt a machine learning approach to identify high-risk online job adverts which may potentially lead to labour exploitation. These techniques enable the identification of relevant linguistic patterns indicative of high-risk adverts. Leveraging these features, a predictive model is trained using a random forest classifier to differentiate high-risk adverts from others. Through this approach, we aim to facilitate timely intervention to protect vulnerable job seekers from exploitation.

While prior research **mdpiAutomaticDetection** has shown that it is possible to identify authentic job advertisements from those that are deceptive and 'fake' , identifying jobs that may potentially lead to exploitation is entirely different. While adverts aiming to scam or ensnare someone into labour exploitation are both deceptive, the incentives of the person posting the advert differ significantly Organisation 2016.

The intention of a scammer is to make the advertisement appear realistic only to the extent necessary to extract valuable information from the applicant, such as phone numbers, financial details, social security information, or login credentials. The scammer aims for the content of the advertisement to appear authentic enough to blend in with legitimate postings, deceiving applicants into divulging sensitive information.

Conversely, an advertisement designed to entice an applicant into exploitative circumstances may offer additional incentives to apply, such as enticingly high salaries or promises of exceptional working conditions, in order to attract vulnerable individuals. In this scenario, the deceiver intends for the advertisement to appear convincing enough not only to prompt application but also to compel the applicant to accept the role. However, the advertised position might not actually exist, or the working conditions described in the listing could be misleading Ada Volodko 2019.

Further complicating matters is the possibility that labour exploitation of the worker may occur at a later stage, even if the job initially appears as advertised. The difficulty in this undertaking lies in identifying advertisements that can definitively be ruled out as non-exploitative. Identifying whether a job advertisement has resulted in labour exploitation is challenging without direct confirmation from the worker or verification from a trusted advertiser. Inspection alone may

not suffice to determine whether a job advertisement has led to exploitation, as such instances often require firsthand reports from affected individuals or validation from reputable sources.

We set out to apply data science techniques to see if applying such methods could successfully distinguish exploitative advertisements from others. Additionally we wanted to uncover any characteristic features of exploitative ads

Our aim to identify such deceptive advertisements is an example of 'positive unknown' (PU) learning, a type of statistical learning problem where instead of negative instances, we encounter a mix of advertisements with unknown outcomes.

Our approach is to treat the advertisements with unknown outcomes as if they were negative and to use the data in this way to train a classification model.

The dataset available for this project consisted of adverts posted on Facebook and were in image form. The sample of adverts available for this project were sourced through collaboration with a non-profit organisation – these advertisements are known to have resulted in instances of labour exploitation. Initially, the dataset was small in size, prompting the collection of additional data from a a Facebook group resembling the original platform.

The primary aim was to balance the dataset, ensuring equal representation of exploitative advertisements and those with uncertain outcomes. However, there were concerns regarding the inclusion of advertisements with unknown outcomes. Seeking and including additional advertisements is a process with diminishing returns, especially if this results in the number of unknown samples being greater than the number of positive ones. There are several reasons for this, the first is that by adding unknown samples the underlying class distribution is obscured. Although we are treating the unknown samples as having negative labels, we don't know which of them are true positives or true negatives.

Finally, the more unknown samples that are included in the dataset, the greater the risk of introducing irrelevant features or noise, leading to overfitting issues when training classification models.

Text was extracted from images of the advertisements using easyOCR, a Python package designed for optical character recognition (OCR) tasks. The script extracts text recognised in other languages, however only the English samples are used to construct the final dataset.

Feature	Description	Data Type
word_len	Advert length in words	Integer
char_len	Advert length in words	Integer
phone_number	Phone number as an integer	Integer
mention_visa	Number of times the word 'visa' is mentioned.	Integer
mention_free	Number of times 'free' is mentioned.	Integer
mention_whatsapp	Number of times 'whatsapp' is mentioned.	Integer
at_symbol	Number of times the '@' symbol is mentioned.	Integer
hash_symbol	Number of times the '#' symbol is mentioned.	Integer
gender_requirement	Number of times gender is referred to.	Integer
application_requirement	No. of mentions of an application task	Integer
dm_request	Number of times advert references any informal contact	Integer
urgency	Number of times there is mention of urgency in the advert	Integer
email	Whether or not the advert contains an email address	BoolInt
conjunction	Total number of conjunctions in the advert	Integer
adverb	Total number of adverbs in the advert	Integer
preposition	Total number of prepositions in the advert	Integer
noun	Total number of nouns in the advert	Integer
adjective	Total number of adjectives in the advert	Integer
polarity	Positive or negative sentiment	Integer
subjectivity	Extent to which personal opinions are expressed.	Integer
professions_n	Vectors representing different professions mentioned	Float
profession_count	Total number of different professions mentioned	Integer
target	Whether the advert led to exploitation or has an unknown outcome	BoolInt

Table 1: Table of all features extracted to create a dataset from job advertisements.

The extracted text is then processed in various ways using text mining as well as sentiment analysis, parts-of-speech tagging and word embeddings using Word2Vec. A dataset is constructed consisted of features that are either integers or floating point numbers for easy input to a series of classification models. Table 1describes all of the extracted features.

The constructed dataset consists of 94 instances and 34 features. The entire dataset was randomised and split as follows to allow for training of a classification model:

- Training set: 64
- Validation set: 16
- Test Set: 20

A Random Forest classifier was trained on the training set. GridSearchCV was used to validate the classification model using a five fold split, the same package was also used to identify which

combination of parameters result in the best performance. The optimised model was then evaluated on the validation set, showing an improvement compared to the training set. Finally, the optimised model was evaluated on the training set. The optimised, validated model has a mean accuracy of 70% - it correctly identifies 77% of all positive results in the data and classifies 68% of all unlabelled instances as negative, on average. The quoted accuracy scores were obtained by training the classifier one hundred times and taking the mean of the accuracy scores reported in every run of the model.

A Naïve Bayes Classifier was also trained on the same dataset, however this model didn't perform as well as the Random Forest Classifier. This could be because the variables weren't truly independent of each other - this is a basic assumption of the Naïve Bayes Classifier. To account for this, principle components analysis (PCA) was used to reduce the dimensionality of the dataset and also project the data onto features that are independent. Despite these attempts to improve model accuracy, the random forest classifier still outperformed the Naïve Bayes Classifier as shown in 3.

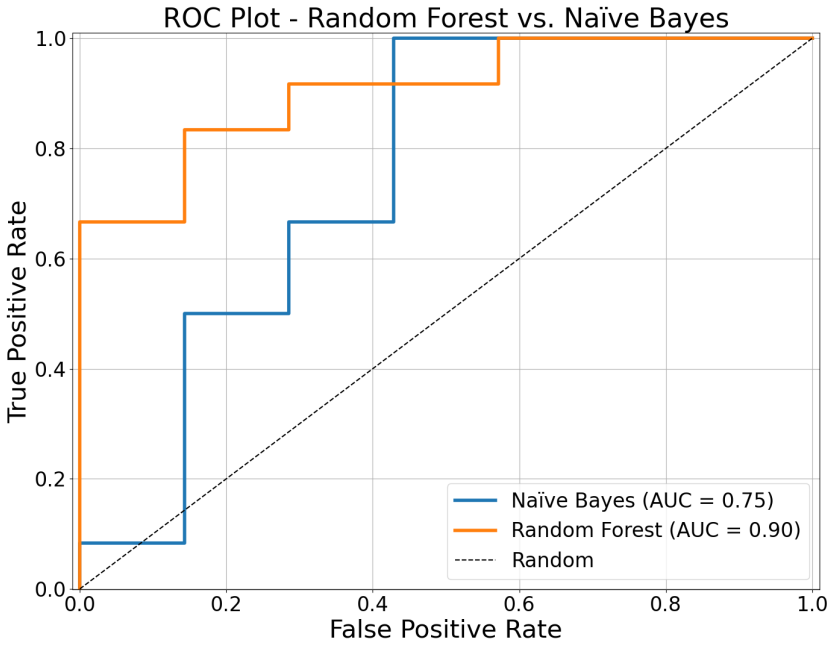

Figure 1: An ROC curve showing the performance of the random forest classification model compared to a Naïve Bayes Classifier.

An ROC curve was generated to illustrate the model's performance, as depicted in figure 3. The line representing the random forest model demonstrates that at various thresholds set by indi-

vidual trees within the model, at least 60% of positive results were accurately classified. Notably, the performance of the random forest model surpasses that of a previous classification attempt employing a naive Bayes classifier, and its curve exhibits significantly higher performance than what would be expected by random chance alone.

Analysing the importance of each feature to the classification model revealed the aspects of the exploitative adverts had the greatest influence over the decision made to classify an advert as exploitative or not.

A notable observation is the model's tendency to classify a significant proportion of 'negative' results as potentially exploitative as evidenced by a high false positive rate. This could indicate that the model has detected deceptive features in the 'unknown' data, as intended. However, further investigation is required to determine a cause for this discrepancy.

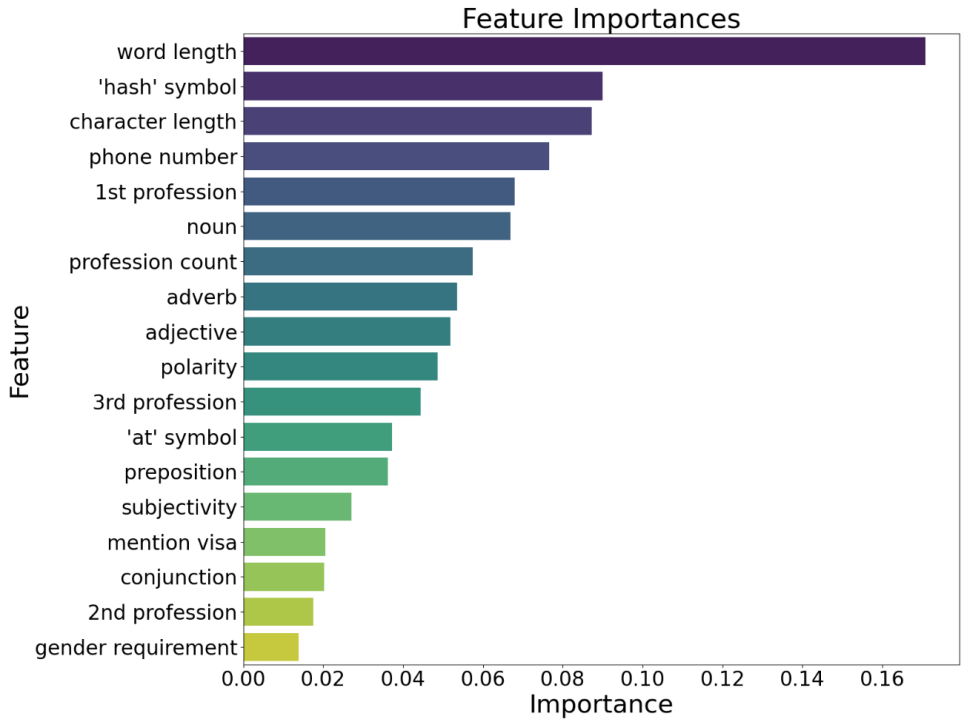

Figure 2: A bar chart showing the influence of each feature on classifications made by the random forest model.

Using random forests, it is possible to analyse the influence each feature has on the classification of an advertisement as deceptive or not. Figure 2 shows the advertisement features in decreasing importance to the decisions made by the classifier, word length is the attribute with the most influence over the classification. It is worth noting that none of the individual features have an exceptionally high influence over the classification decision, indicating a balanced model without

signs of overfitting or bias.

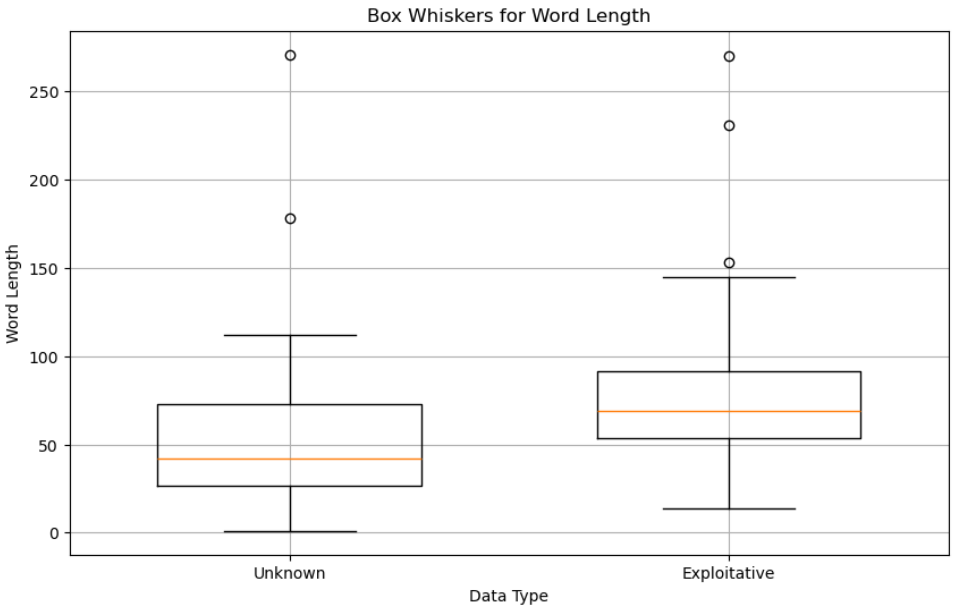

Figure 3: Box and whisker plots comparing the distribution of ad length in words for advertisements with unknown outcomes (left) and exploitative advertisements (right).

It is worth noting however, that without additional data it is difficult to establish if the classification models are truly distinguishing the positive samples from the unknown samples or if the classifier is simply modelling differences in the data used to construct the dataset. Figure 3 shows box and whisker plots showing the distribution of advertisement word lengths in the original sample and the added advertisements. When a feature is recognised as influential in the classifier's decisions, we anticipate this will reveal a causal factor. How can we be certain that this outcome doesn't simply reflect disparities between the two datasets?

0.1 Future Work and Recommendations

- Investigate the difference in text extracted using OCR compared to other image processing methods like convolutional neural networks (CNNs).
- Investigate further the cause of the high false positive rate.
- Continue research to enhance feature set and model robustness.
- Explore other methods for dealing with a lack of negative samples, e.g. anomaly detection methods as applied to spam and malware detection.

- Extend the work by employing Large Language Models (LLMS).
- The ocr extraction script is able to extract text in other languages, translate these and combine/compare these extracts to the English samples. Does this result in improved performance.
- Collaborate with stakeholders to deploy intervention strategies based on model predictions.

Bibliography

Ada Volodko, Ella Cockbain Bennett Kleinberg (2019). ""Spotting the signs" of trafficking recruitment online: exploring the characteristics of advertisements targeted at migrant job-seekers - Trends in Organized Crime — link.springer.com". In: [Accessed 27-03-2024].

Organisation, International Labour (2016). *Deceptive recruitment and coercion — ilo.org*. [Accessed 27-03-2024].

– (2022). "Global Estimates of Modern Slavery: Forced Labour and Forced Marriage". In: [Accessed 27-03-2024].

References

Ada Volodko, Ella Cockbain Bennett Kleinberg (2019). ""Spotting the signs" of trafficking recruitment online: exploring the characteristics of advertisements targeted at migrant job-seekers - Trends in Organized Crime — link.springer.com". In: [Accessed 27-03-2024].

Organisation, International Labour (2016). *Deceptive recruitment and coercion — ilo.org*. [Accessed 27-03-2024].

– (2022). "Global Estimates of Modern Slavery: Forced Labour and Forced Marriage". In: [Accessed 27-03-2024].