Recruitment Scam Detection

Tanaya Kavathekar, Swetha Kalla, Jyoti Sharma, Tanvi Hindwan

Data Science Department Columbian College of Arts and Science

George Washington University, Washington DC

Table of Contents

Introduction:

NLP applications work closely with text analysis and text mining. With the help of these applications, we can extract important information that can resolve our various real-life problems like online recruitment scam. The issue of financial loss, identity theft, credibility loss, and demotivation suffered by the individuals who get trapped in these online recruitment frauds is of paramount importance. Therefore, it is required to analyze whether the job post is legit or not. Working on this kind of real-life problem encourages us to do precise text analysis, handle imbalanced data, and to develop a model that can learn to distinguish between legit and fraud job posts.

The present research aims to identify fraud recruitment posts by extracting information from entities such as job descriptions, benefits, requirements, and company profiles. The project attempts to leverage language modeling techniques such as n-gram to identify high-frequency words in the legit and fraud job postings, sentence clustering to form distinctive clusters pertaining to fraudulent and legitimate posts, and classification techniques to distinguish between legit and fraud posts.

The dataset used in research is obtained from Kaggle with 17,014 legitimate and 866 fraudulent job ads, containing 18 variables such as job description, requirements, benefits, and company profile, etc. The data consists of both textual information and meta-information about the jobs.

Background:

For the initial stages of analysis, the draft "Language models for information retrieval" by Stanford NLP group was referred. The draft explained the concept of various language models which can be applied for a query like information retrieval. And for the project's analysis, language modeling using n-gram (3-grams and 4-grams) is used to build probability distribution over a sequence of terms.

For the classification, the choice of the models was based on "Text Categorization with Support Vector Machine", where the author has detailed text categorization using Support Vector Machines. SVM model performs robustly on text classification problems without the need for eliminating hyper-parameter tuning.

Scope:

The dataset was first cleaned using various preprocessing steps like HTML tags removal, punctuations, stopwords removal, email id, and trademarks. Further, tokenization and lemmatization techniques were used to analyze the frequency distribution of tokens for fraud and

legit posts. While performing exploratory data analysis it was found that the industry Information Technology and Services (Figure 1) has the highest number of job posts. Therefore it was first attempted to narrow down the insight into the IT industry to analyze the most frequently occurring fraud and legit job description terms. However, the outcomes were not that persuasive and were found overlapping with the legit job postings.

For a deeper insight into the pattern of fraud and legit terms, n-gram models were used to build probability distribution over a sequence of terms. Instead of calculating single word count, the probability of a word in legit and fraud postings was counted in the form of trigrams and four-gram. And it was found that certain sequences of words in n-grams are actually from the fraudulent job posts for example ‘aker solution global’, ‘participate referral bonus’ etc.

In order to analyze the variation in lengths and punctuations within the job posts, frequency distribution graphs are used. Figure 2 and Figure 3 show that legit and fraud posts have nearly the same distribution. Hence there is no relationship between length and punctuation with two categories of job posts.

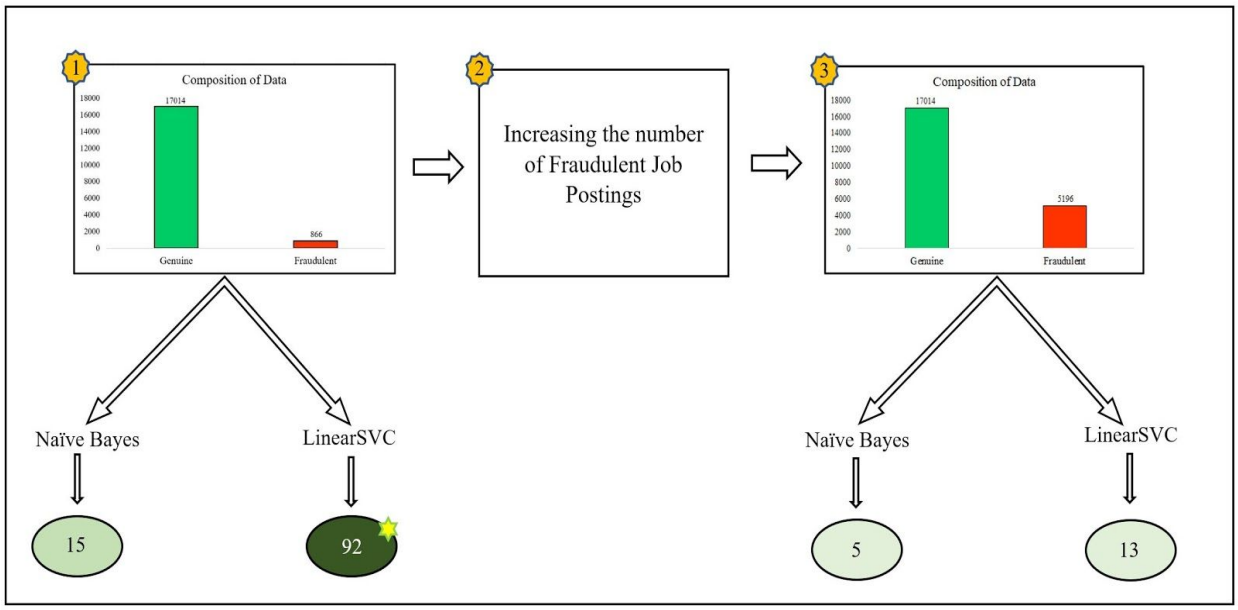
Another aspect of this project was to identify fraudulent sentences within the job posting by using sentence level clustering. The available raw data was used for this analysis, without augmentation, as this was an unsupervised learning experiment. The clusters were formed within the industry of Oil and Energy as it had a good balance of fraud(178) and legit(109) posts. In order to represent texts as a vector, TF-IDF vectorization was performed. TF-IDF measures the importance of a word within a job posting by calculating a composite score for each word[2]. On these word vectors, a minibatch Kmeans clustering technique was applied with a batch size of 2048. Figure 3 shows the SSE plot for 20 clusters. It can be seen that there is a drop in error at K=16, hence it was decided to form 16 clusters. Table 1 shows the number of fraud and legit post jobs within clusters. As clusters 3 and 6 have only fraud sentences it can be concluded that sentences mapped to these clusters are a fraud.

Later, attempts were made to perform modeling to see if the algorithms can classify fake postings from the fraudulent postings. For this, we employed Naive Bayes and LinearSVC classifiers. One of the difficulties faced during modeling is the nature of the dataset. Since the number of fraudulent postings in the dataset is way less than the genuine postings, the modeling results were not reliable.

To overcome this problem, techniques like back translation, random deletion were investigated to balance the dataset. Since back translation involved a paid version of Google Translate API, we chose to randomly shuffle and delete some of the text to create new fraudulent postings.

To test these techniques, the algorithms were implemented before and after the augmentation. Since the goal here is to correctly classify the fraudulent postings from the fake postings, the evaluation of the models is done based on how well they were able to classify fake postings.

Upon splitting the data, the test set consisted of 3423 genuine postings and 153 fraudulent postings. The idea is to see how many of these 153 postings the models can classify correctly before and after augmentation, to see if augmentation has truly helped in better classification.



From the above figure, we can see that the Linear SVC model has better performance than Naïve Bayes and was able to classify the fraudulent postings better on the raw data rather than the augmented data. If anything, augmenting the data has proven to be harmful in this case.

Outcome:

In the sentence clustering exercise, it was observed that out of 16 clusters 10 clusters had sentences from both legit and fraud posts and 2 clusters had only fraud posts while 4 clusters had legit posts. As this technique only considers the importance of words within documents, these clusters are not a good representative of the fraudulent sentences. Hence it is concluded that sentence clustering on this dataset does not provide concrete results.

Further, it is intended to address the issue of imbalance data in NLP projects by performing and evaluating different augmentation techniques on the dataset. But this strategy has proven to be fatal in this case. This can be attributed to poor augmentation and sampling strategies, or partly because feature engineering was not performed on the cleaned dataset, which resulted in deceptive outcomes.

As for the modeling results, the LinearSVC model has performed better than Naive Bayes. This can be due to the reason that most of the text classification problems are linearly separable [5, 4] and works well on high-dimensional input spaces.

Challenges:

The problem faced with implementing back translation for augmentation was that the translation was implemented using the TextBlob library, which internally uses Google Translate API with rate-limited calling, resulting in errors like 'Too many requests'.

Another serious issue experienced in the augmentation exercise was that the produced content was incredibly like the original content since the information is exceptionally specialized, (for example, Chemical, Electrician, and so forth).

Additionally, job posts metadata such as IP address, location, and company electronic verification number play a vital role in recognizing fraudulent job posts. It is very difficult to distinguish between the fraud and legit job posts using only text mining as both categories of posts are alike.

Future Scope:

In the project, there is a scope of improvement for better interpretations. Thus, for future work, a similar analysis can be performed with larger and latest datasets to get better and more reliable outcomes.

Even for performing data augmentation, a paid version of Google translate API or better frameworks can be utilized which can make meaningful results for performing classification.

Also, different vectorization and clustering techniques can be implemented which could reflect more characteristics of the fake job posting.

References

1. Language models for information retrieval. (2009, April 1). Retrieved from https://nlp.stanford.edu/IR-book/pdf/12lmodel.pdf

2. (n.d.). Retrieved from https://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

3. V.R, A. (2020, January 23). Recruitment Scam. Retrieved from https://www.kaggle.com/amruthjithrajvr/recruitment-scam

4. Joachims, T. (n.d.). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Retrieved from https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

5. KOWALCZYKI, A. (2018, November 14). Linear Kernel: Why is it recommended for text classification? Retrieved from https://www.svm-tutorial.com/2014/10/svm-linear-kernel-good-text-classification/

Acronyms

HTML - Hypertext Markup Language

TF-IDF - Term Frequency Inverse Document Frequency

SSE - Sum of Squared Error defined as the sum of the squared distance between the centroid and each member of the cluster.

SVC - Support Vector Classifier

Appendix

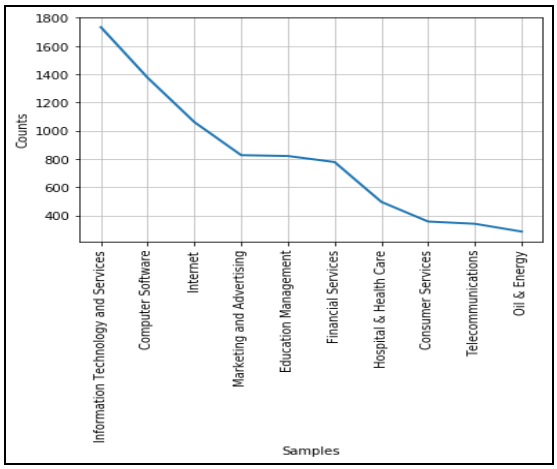


Figure 1. Analysing Industry with higher number of job posts

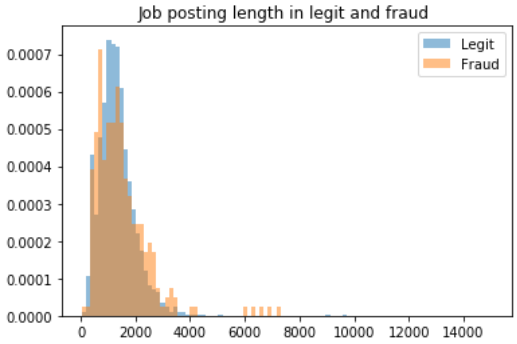


Figure 2 Length of the legit and fraud job posts

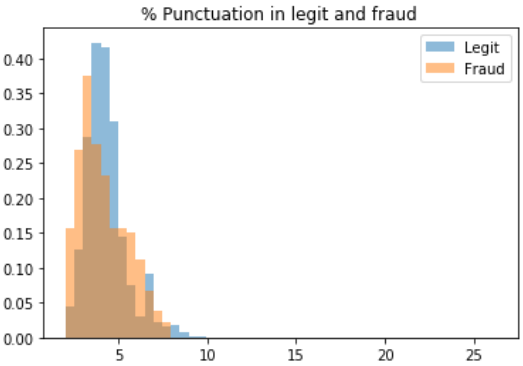


Figure 3 Percentage of punctuation of the legit and fraud job posts

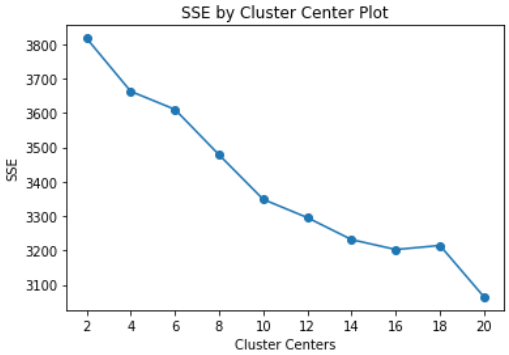


Figure 4 SSE by cluster centers

Table 1 showing the distribution of legit and fraud posts within the clusters
Red cells cluster have only fraud posts, green cell clusters have legit posts

Clusters	Legit posts	Fraudulent posts
1	817	424
2	137	230
3	0	63
4	29	117
5	107	0
6	0	68
7	79	104
8	263	45
9	88	18
10	58	0
11	283	319
12	106	0
13	107	0
14	93	34
15	174	138
16	107	0