FAKE JOB POSTINGS DETECTION

Freddy Chen
Keani Schuller
Kelly Kao
Ko-Jen Wang
Xinran Yu

TABLE OF CONTENTS

01 INTRODUCTION

02 EDA

03 MODEL BUILDING

04 RESULTS

05 CONCLUSION

01

INTRODUCTION

TOO GOOD TO BE TRUE?

[redacted] • 3rd+
Airport Shuttle Drivers
3mo •

Primary duties include picking up passengers providing on time shuttle service assisting passengers by answering question, providing directions or attending to their Various other needs. This includes helping paengers with inquiries about delayed or lost baggage.

$2,000-10,000/Week

1 • 1 comment

Salary seems good. Where can I apply?

Work From Home Position

Reply | Reply All | Forward | ...

Thu 12/10/2020 10:36 AM

ES Evan Ston
To

You forwarded this message on 12/10/2020 10:39 A.M.

Hello,
I am a staff in the college, a professor of Medicine shared me a link for students who might be interested in a PAID UNICEF PART TIME POSITION job to make up to $500 weekly,

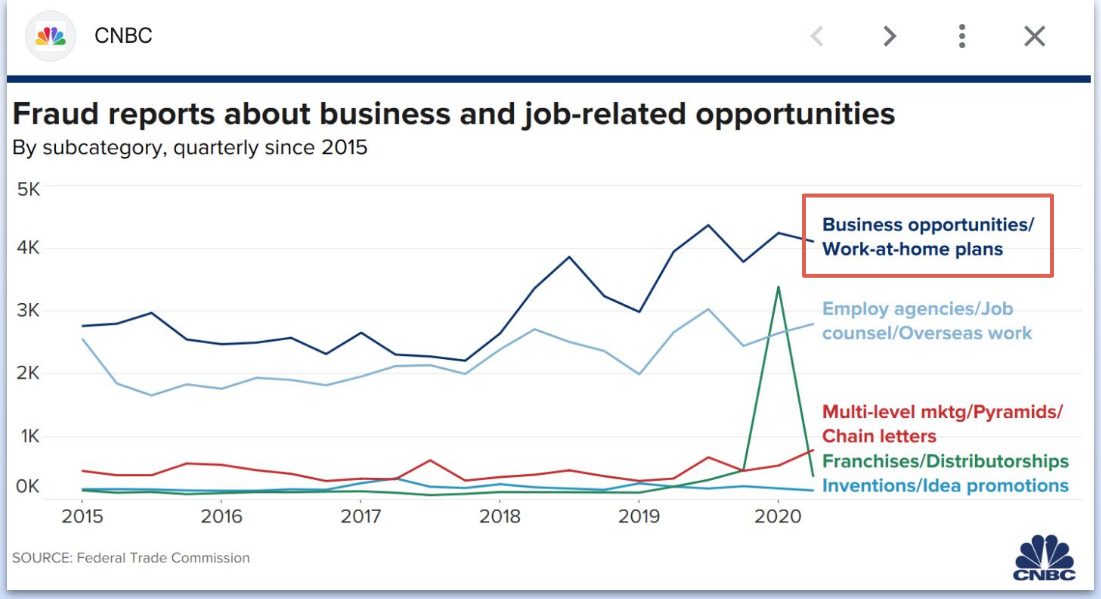
Email for More info - freemanjohnny6s66@outlook.com

NOTE: This is strictly Work From Home Position

Do have a good day,
Thank you

I can work from home forever?

FAKE JOB POSTING TREND



If you're looking for a job on LinkedIn, there are a few things you can do to protect yourself from fake job postings:

- Be wary of job postings that offer unrealistic salaries or benefits.
- Do a Google search for the company name and the job title to see if the posting is legitimate.
- Be suspicious of job postings that ask you to provide personal information before you've even been interviewed.
- Never wire money to a company or individual you don't know.

THIS IS A CONCERNING ISSUE...

News / Local News

Latest scam sees malware spread through fake job postings on LinkedIn

The Whig-Standard

Published Nov 21, 2023 • Last updated Nov 22, 2023 • 1 minute read

Join the conversation

Police are warning of malware being spread through fake job postings on LinkedIn.

FBI EL Paso

Twitter Facebook Email

April 21, 2021

FBI Warns Cyber Criminals Are Using Fake Job Listings to Target Applicants' Personally Identifiable Information

Press release available in both English and Spanish.

Scammers advertise jobs the same way legitimate employers do—online (in ads, on job sites, college employment sites, and social media), in newspapers, and sometimes on TV and radio. They promise you a job, but what they want is your money and your personal information.

Fake Job or Employment Scams occur when criminal actors deceive victims into believing they have a job or a potential job. Criminals leverage their position as "employers" to persuade victims to provide them with personally identifiable information (PII), become unwitting money mules, or to send them money.

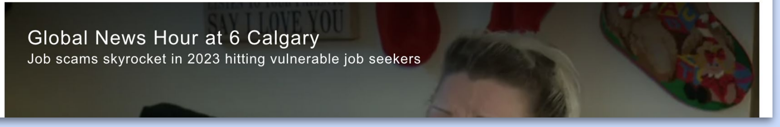
Fake Job Scams have existed for a long time but technology has made this scam easier and more lucrative. Cyber criminals now pose as legitimate employers by spoofing company websites and posting fake job openings on popular online job boards. They conduct false interviews with unsuspecting applicant victims, then request PII and/or money from these individuals. The PII can be used for any number of nefarious purposes, including taking over the victims' accounts, opening new financial accounts, or using the victims' identity for another deception scam

CONSUMER

Job scams skyrocket in 2023, targeting vulnerable employment seekers

By Tomasia DaSilva • Global News

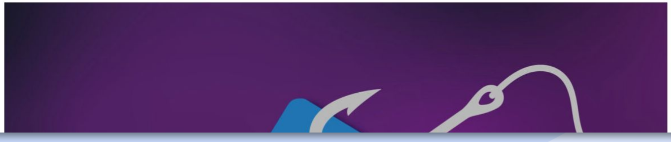
Posted December 12, 2023 7:59 pm · Updated December 13, 2023 12:18 pm · 3 min read



Home » Security Bloggers Network » LinkedIn Fakes: The Rise of Spoof Profiles Threatening Brand Reputation

LinkedIn Fakes: The Rise of Spoof Profiles Threatening Brand Reputation

by Sam Bakken on August 24, 2023



PROJECT GOAL

The project aims to identify the fake job postings through text analysis and classify them into real and fake groups.

Datasets

From University of the Aegean

Laboratory of Information & Communication Systems Security

- 18K job postings

job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function	fraudulent
1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	NaN	0	1	0	Other	Internship	NaN	NaN	Marketing	0
2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...	0	1	0	Full-time	Not Applicable	NaN	Marketing and Advertising	Customer Service	0
3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...	NaN	0	1	0	NaN	NaN	NaN	NaN	NaN	0
								Our									

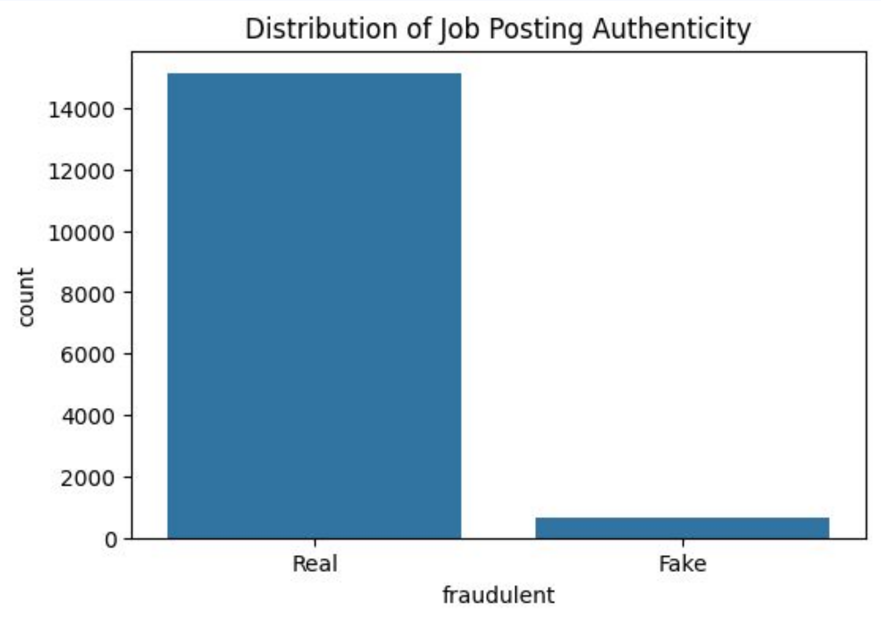
02

EXPLORATORY DATA ANALYSIS

HISTOGRAMS

Job Posting Authenticity

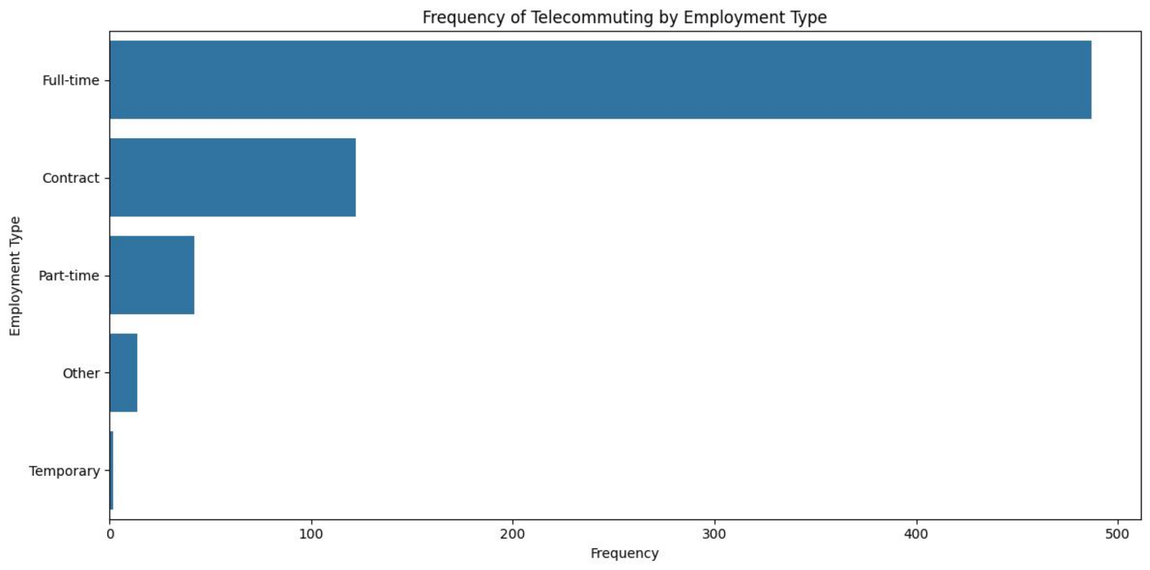
- Comparison of the number of real job postings to the number of fake job postings
- There are significantly more real postings versus fake postings at around 90% Real and 10% Fake
- The dataset is therefore imbalanced
- This however reflects reality, since there aren't usually a large amount of fake job postings on recruiting sites



HISTOGRAMS

Telecommuting Frequency

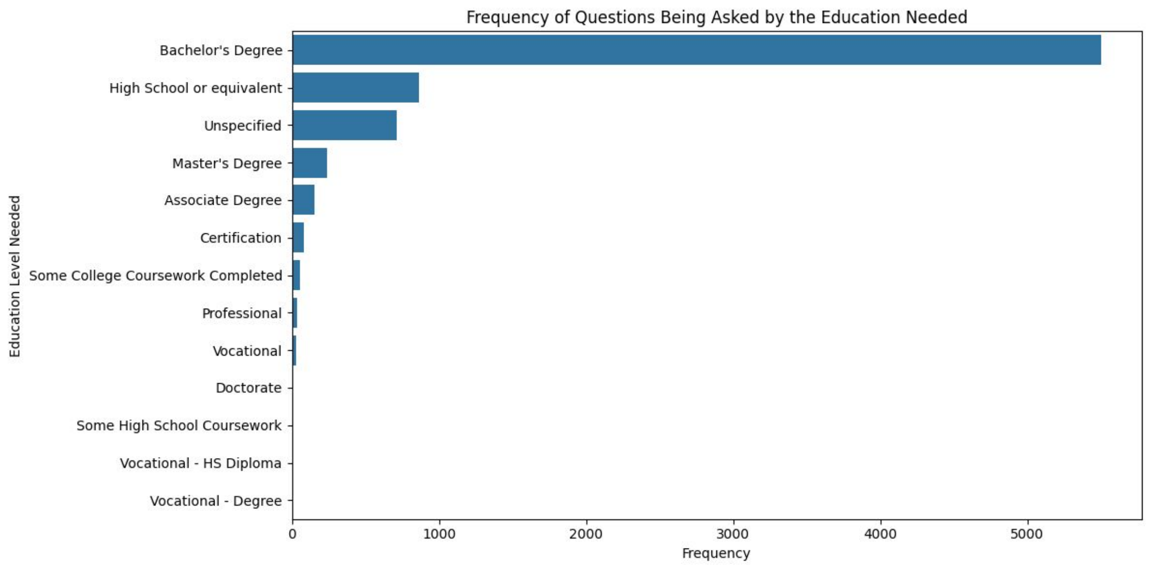
- Comparing the frequency of telecommuting by employment type
- Full-time work is the most likely to include telecommuting by a large margin
- The second most likely is contract



HISTOGRAMS

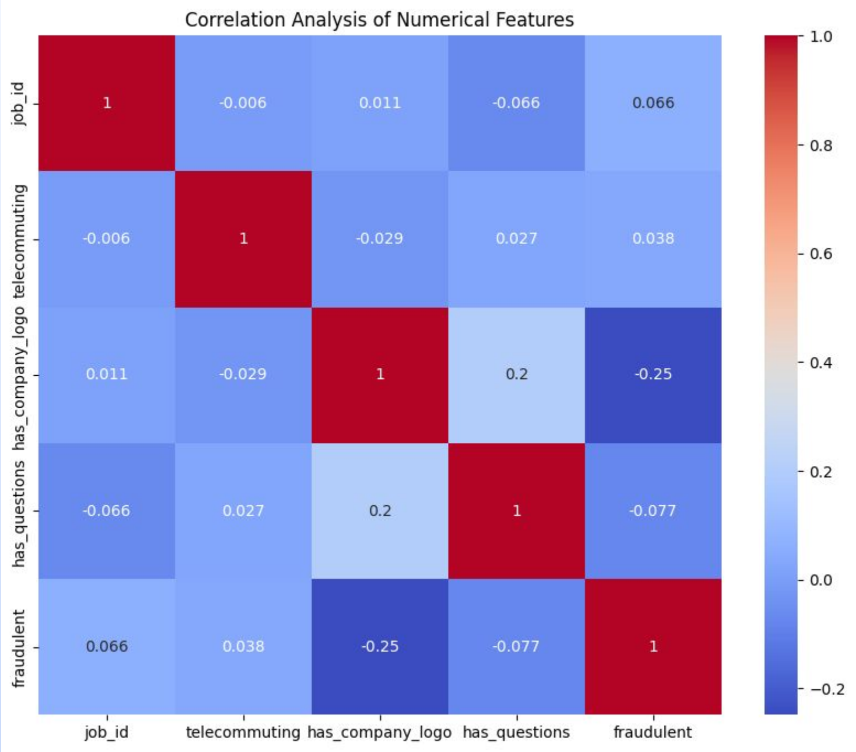
Questions Frequency

- Comparing the frequency of questions by education level required
- Employment requiring a bachelor’s degree has the highest frequency of questions by a very large margin



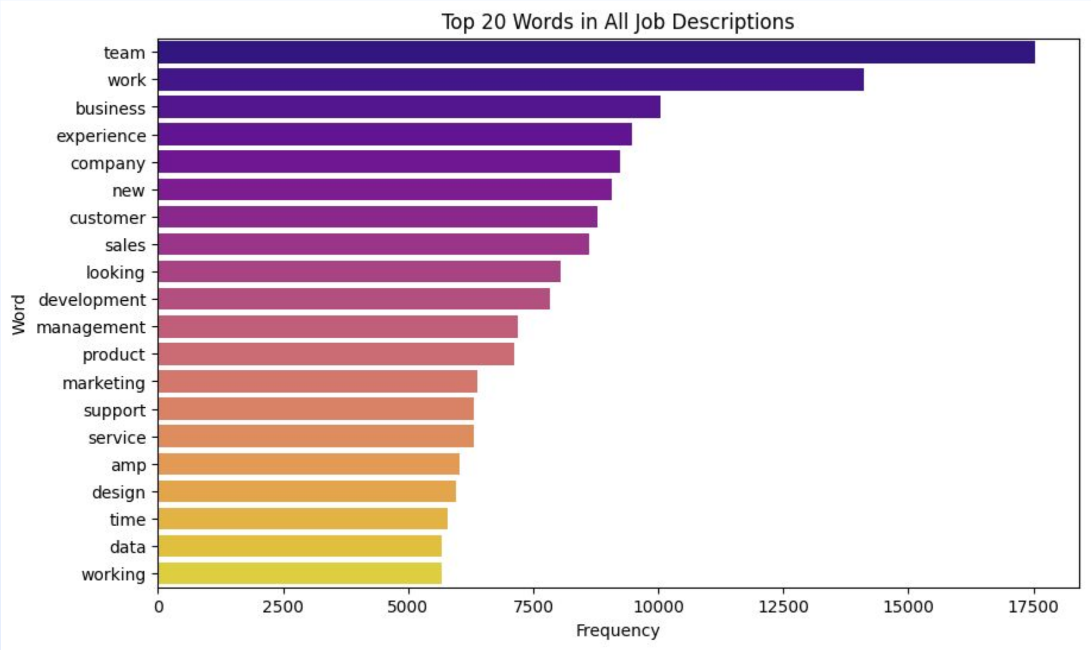
CORRELATION MATRIX

- Correlation matrix to visualize the correlations between all numerical predictors
- Overall, the correlations are very low from -0.25 to 0.2
- The strongest correlation is between the company having a logo and the target variable fraudulent, which is a binary indicating whether the job posting is real or not
- The second strongest is between the company having a logo and the number of questions



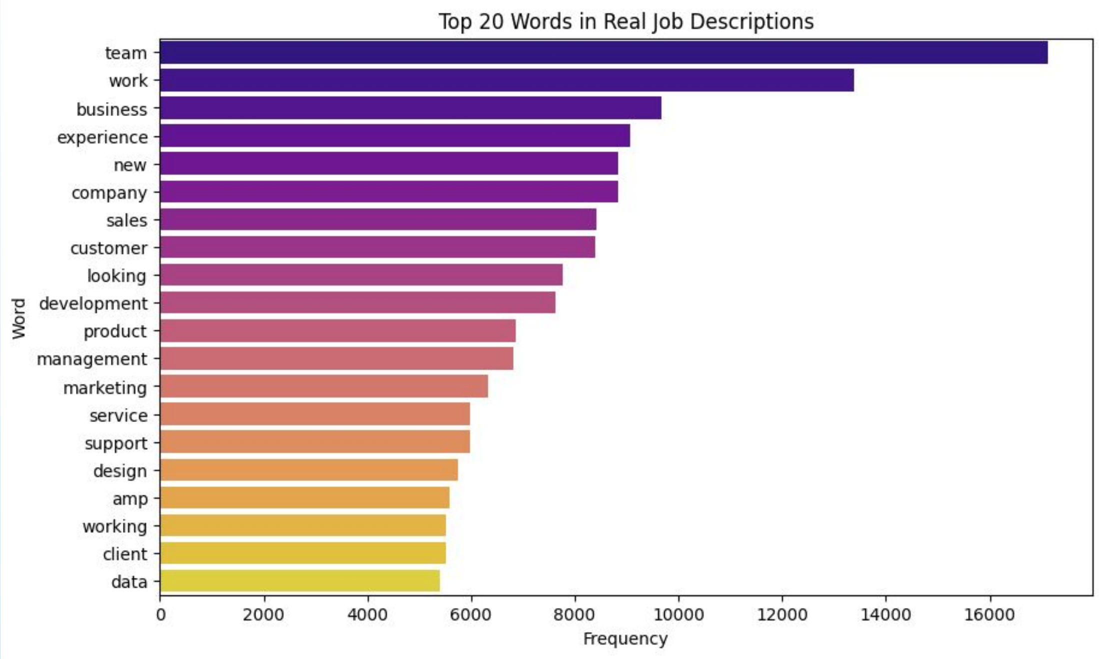
TOP 20 JOB DESCRIPTIONS WORDS

- Histogram demonstrating the top 20 words in all the descriptions of the job postings
- The most frequent word overall was “Team”
- This is followed by “Work”, “Business” and “Experience”
- It also includes words like “Sales”, “Customer”, and “Product”



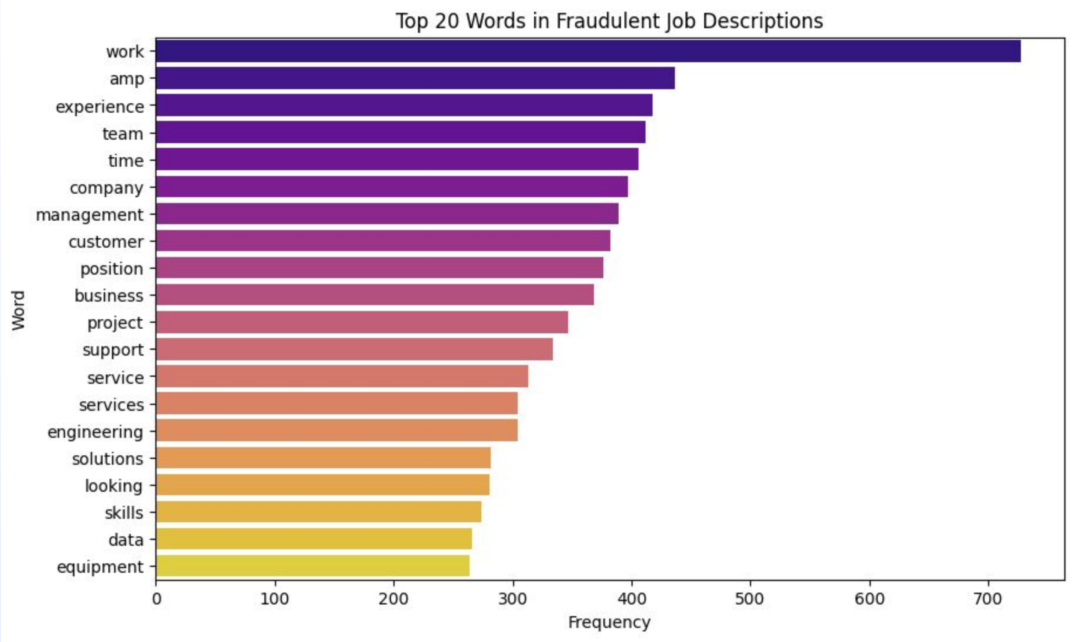
TOP 20 WORDS FOR REAL JOBS

- Histogram demonstrating the top 20 words in Real job descriptions
- The most frequent word overall was “Team”
- Real job postings include words like “Sales”, “Customer”, “Development”, “Management”, and “Client”



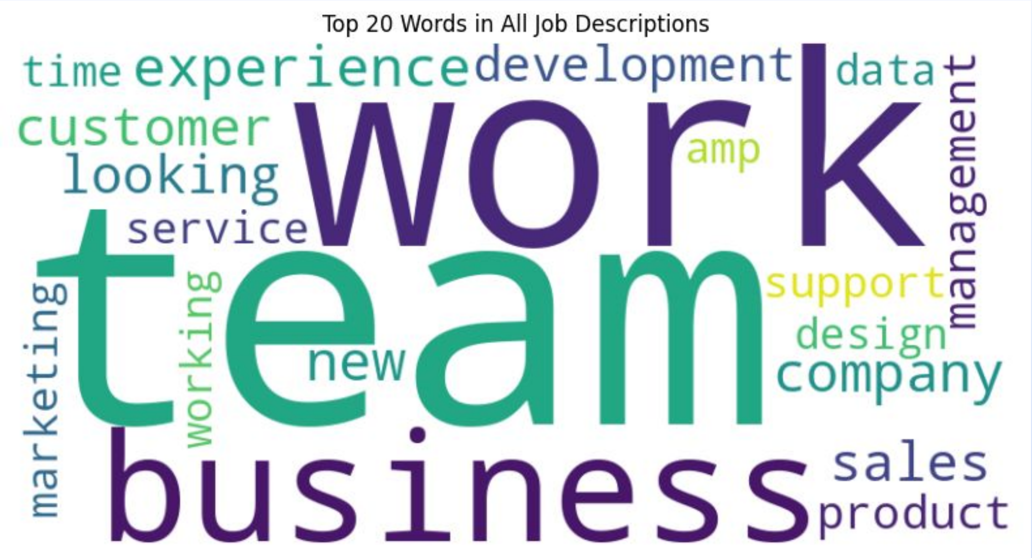
TOP 20 WORDS FOR FAKE JOBS

- Histogram demonstrating the top 20 words in Fake job descriptions
- The most frequent word overall was “Work”
- This is followed by “Amp” which seems like a frequent typo
- Fake job postings include words like “Experience”, “Time”, “Company”, “Support”, and “Solutions”
- A curious observation is that there are many mentions of engineering in fake job postings



WORD CLOUD

- Word cloud demonstrating the top 20 words in all job descriptions
- This was made to be able to better visualize the top words
- The results are the same as that in the histogram showed before, but was easier to read in this format



03

MODEL BUILDING

EXPERIMENTS

12 Combinations:

- Preprocessing w/wo Lemmatization
- TF-IDF Vectorizer: (1, 1), (1, 3), (2, 2)
- Models: MultinomialNB/ SVM

MODELS SELECTED

MultinomialNB

- Suitable for Word Frequency Analysis
- Efficient and Scalable

SVM

- Diverse Kernel Methods
- High-Dimensional Performance

Different parameter tuning is applied for different models.

- Alpha

- C (controlling outliers)
- Kernels

MODEL BUILDING

`preprocess_text(text, lemmatize=False)`

- Preprocess the input text data by removing stop words, non-alphabetic tokens, and optionally lemmatizing

`run_experiment(dataframe, text_column, label_column, experiment_config, grid_search_configs)`

- Configure a machine learning pipeline with `TfidfVectorizer` and a classifier as specified in the experiment configuration
- Use grid search to find the best hyperparameters for the pipeline based on cross-validation.
- Retrain the model with the best hyperparameters on the training data.
- Evaluate the model on the test data using accuracy, recall, precision, and generate a confusion matrix.
- Store and return the results of the experiment and the confusion matrices for analysis.

```
experiment_configs = [
  {
    'preprocessing':,
    vectorizer_condition:,
    classifier:
  },
]
```

```
grid_search_configs = [
  {
    'classifier_name':,
    'classifier':,
    'param_grid':{
    }
  },
]
```

04

RESULTS

MODEL RESULTS

There is about a 20% difference in the recall rates between MultinomialNB and SVC.

Preprocessing	Vectorizer Condition	Classifier	Accuracy	Recall	Precision
Yes	(1,1)	MultinomialNB	0.973	0.454	0.855
Yes	(1,3)	MultinomialNB	0.974	0.473	0.891
Yes	(2, 2)	MultinomialNB	0.980	0.565	0.975
No	(1,1)	MultinomialNB	0.974	0.483	0.862
No	(1,3)	MultinomialNB	0.975	0.483	0.885
No	(2, 2)	MultinomialNB	0.981	0.575	0.967
Yes	(1,1)	SVC	0.983	0.720	0.876
Yes	**(1,3)**	**SVC**	**0.987**	**0.734**	**0.956**
Yes	(2, 2)	SVC	0.986	0.691	0.973
No	(1,1)	SVC	0.985	0.749	0.896
No	**(1,3)**	**SVC**	**0.987**	**0.734**	**0.956**
No	(2, 2)	SVC	0.986	[illegible]	[illegible]

The impact of lemmatization is not significant.

While sacrificing some precision, the model can achieve a recall rate of around 73.4%.

MODEL RESULTS – Confusion Matrix

Preprocessing	Vectorizer Condition	Classifier	TP	FN	FP	TN
Yes	(1,1)	MultinomialNB	94	113	16	4514
Yes	(1,3)	MultinomialNB	98	109	12	4518
Yes	(2, 2)	MultinomialNB	117	90	3	4527
No	(1,1)	MultinomialNB	100	107	16	4514
No	(1,3)	MultinomialNB	100	107	13	4517
No	(2, 2)	MultinomialNB	119	88	4	4526
Yes	(1,1)	SVC	149	58	21	4509
Yes	(1,3)	SVC	152	55	7	4523
Yes	(2, 2)	SVC	143	64	4	4526
No	(1,1)	SVC	155	52	18	4512
No	(1,3)	SVC	152	55	7	4523
No	(2, 2)	SVC	144	63	4	4526

MODEL RESULTS (Accuracy/Recall/Precision)

Preprocessing	Vectorizer Condition	Classifier	Accuracy	Recall	Precision
Yes	(1,1)	MultinomialNB	0.973	0.454	0.855
Yes	(1,3)	MultinomialNB	0.974	0.473	0.891
Yes	(2, 2)	MultinomialNB	0.980	0.565	0.975
No	(1,1)	MultinomialNB	0.974	0.483	0.862
No	(1,3)	MultinomialNB	0.975	0.483	0.885
No	(2, 2)	MultinomialNB	0.981	0.575	0.967
Yes	(1,1)	SVC	0.983	0.720	0.876
Yes	(1,3)	SVC	0.987	0.734	0.956
Yes	(2, 2)	SVC	0.986	0.691	0.973
No	(1,1)	SVC	0.985	0.749	0.896
No	(1,3)	SVC	0.987	0.734	0.956
No	(2, 2)	SVC	0.986	0.696	0.973

MODEL RESULTS (Confusion Matrix)

Preprocessing	Vectorizer Condition	Classifier	TP	FN	FP	TN
Yes	(1,1)	MultinomialNB	4514	16	113	94
Yes	(1,3)	MultinomialNB	4518	12	109	98
Yes	(2, 2)	MultinomialNB	4527	3	90	117
No	(1,1)	MultinomialNB	4514	16	107	100
No	(1,3)	MultinomialNB	4517	13	107	100
No	(2, 2)	MultinomialNB	4526	4	88	119
Yes	(1,1)	SVC	4509	21	58	149
Yes	(1,3)	SVC	4523	7	55	152
Yes	(2, 2)	SVC	4526	4	64	143
No	(1,1)	SVC	4512	18	52	155
No	(1,3)	SVC	4523	7	55	152
No	(2, 2)	SVC	4526	4	63	144

MODEL RESULTS (Best Parameters)

Preprocessing	Vectorizer Condition	Classifier	Best Parameters
Yes	(1,1)	MultinomialNB	{'MultinomialNB__alpha': 0.1, 'tfidf__max_df': 1.0, 'tfidf__max_features': 10000, 'tfidf__min_df': 2, 'tfidf__ngram_range': (1, 1)}
Yes	(1,3)	MultinomialNB	{'MultinomialNB__alpha': 0.1, 'tfidf__max_df': 0.5, 'tfidf__max_features': None, 'tfidf__min_df': 2, 'tfidf__ngram_range': (1, 3)}
Yes	(2, 2)	MultinomialNB	{'MultinomialNB__alpha': 0.1, 'tfidf__max_df': 0.5, 'tfidf__max_features': None, 'tfidf__min_df': 1, 'tfidf__ngram_range': (2, 2)}
No	(1,1)	MultinomialNB	{'MultinomialNB__alpha': 0.1, 'tfidf__max_df': 0.5, 'tfidf__max_features': 10000, 'tfidf__min_df': 1, 'tfidf__ngram_range': (1, 1)}
No	(1,3)	MultinomialNB	{'MultinomialNB__alpha': 0.1, 'tfidf__max_df': 0.5, 'tfidf__max_features': None, 'tfidf__min_df': 2, 'tfidf__ngram_range': (1, 3)}
No	(2, 2)	MultinomialNB	{'MultinomialNB__alpha': 0.1, 'tfidf__max_df': 0.5, 'tfidf__max_features': None, 'tfidf__min_df': 1, 'tfidf__ngram_range': (2, 2)}
Yes	(1,1)	SVC	{'SVC__C': 10, 'SVC__kernel': 'linear', 'tfidf__ngram_range': (1, 1)}
Yes	(1,3)	SVC	{'SVC__C': 10, 'SVC__kernel': 'linear', 'tfidf__ngram_range': (1, 3)}
Yes	(2, 2)	SVC	{'SVC__C': 10, 'SVC__kernel': 'linear', 'tfidf__ngram_range': (2, 2)}
No	(1,1)	SVC	{'SVC__C': 10, 'SVC__kernel': 'linear', 'tfidf__ngram_range': (2, 2)}
No	(1,3)	SVC	{'SVC__C': 10, 'SVC__kernel': 'linear', 'tfidf__ngram_range': (1, 3)}
No	(2, 2)	SVC	{'SVC__C': 10, 'SVC__kernel': 'linear', 'tfidf__ngram_range': (2, 2)}

05

CONCLUSION

BUSINESS IMPACT

- Fake job postings are important due to their impact on individuals' livelihoods and their trust in the digital job market
- Enhances the credibility of job platforms and the companies that rely on them for recruitment
- Protects job seekers from fraud, contributing to a safer and more trustworthy job search environment
- The integrity of the job market is a foundational element for stability and growth in industries
- Essential for maintaining a healthy employment ecosystem and protecting the interests of all stakeholders involved

Thank You!

Q&A