



HOW TO DETECT JOB RECRUITING FRAUD

Team REAL
Danny, Allen, Tommy, Caroline, Mochi

TABLE OF CONTENTS

01



Introduction

02



Preprocessing
&
EDA

03



NN architecture
&
Pre-trained Model

04



Other Findings
&
Conclusions

TABLE OF CONTENTS

01



Introduction

02



Preprocessing
&
EDA

03



NN architecture
&
Pre-trained Model

04



Other Findings
&
Conclusions

HAVE YOU EVER SEEN THIS?

From: [Redacted] <[redacted]@gmail.com>

Subject: JOB POSITION OPPORTUNITY FOR STUDENT

Dear Student,

I'm very happy to inform you about the job opportunity in conjunction with your school (The University of Houston) we got your mail from your school data base. Our reputable company (CiscoSystems Company) is running a student empowerment program. This program is to help the hardworking student to secure a work at home job, this will not stop you from your daily works and your school activities. All you need is jst an hour or two to carry out the job weekly. Your wages will be \$350 USD per week.

Kindly get back to us with your PHONE NUMBER AND PERSONAL EMAIL IF YOU ARE INTERESTED IN THE JOB POSITION

PHONE NUMBER:

PERSONAL EMAIL:

Regards

[Redacted]

Recruiting Manager

CiscoSystems

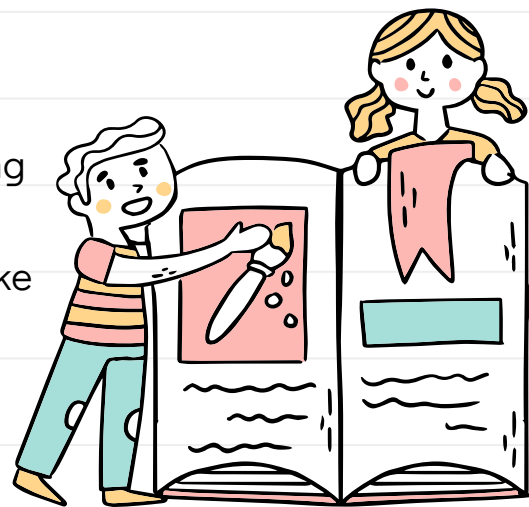
Job Scam!

BACKGROUND

- The problem motivates us is the possibility of **identifying fake job recruiting news**
- People who posted fake job recruiting news may aim at
 - Collecting personal information (even sell it)
 - Providing lucrative job opportunities and asking for money

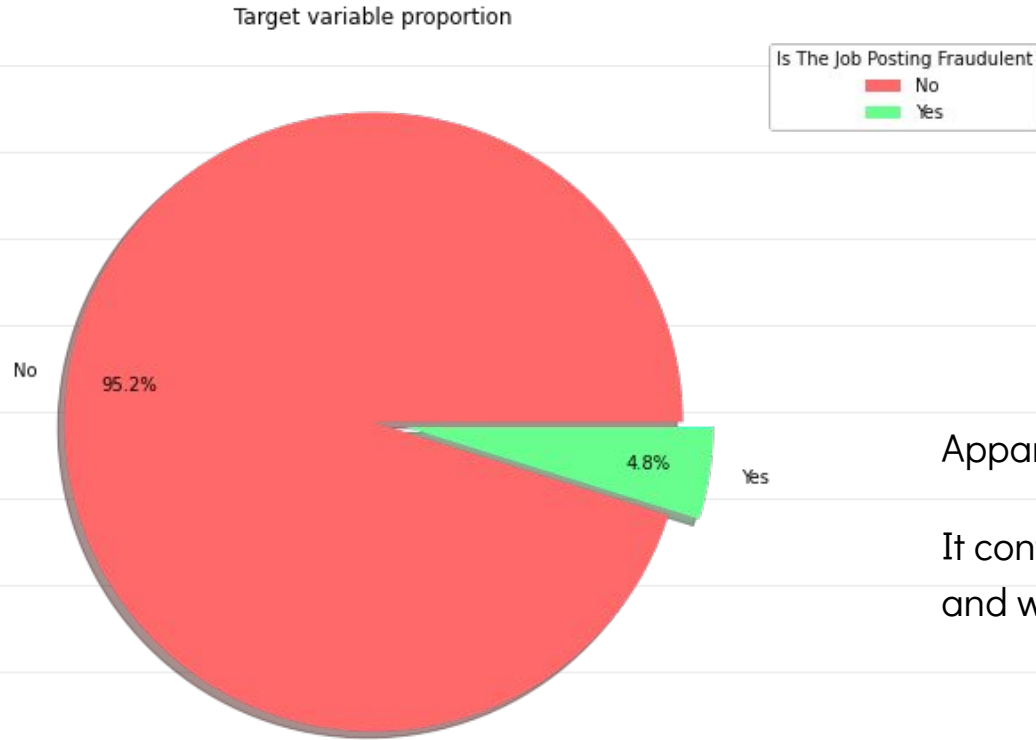
The methods we are considering includes but not limited to:

- Recognition of fake job posting by deep learning (NN) skills, including transferring text into code, model training & testing ect.
- Utilize LIME, sentiment analysis to check the similarities between fake ones



DESCRIPTIVE STATISTICS

- IS THIS DATASET BALANCED?



17,879
postings

866
fake

Apparently, this dataset is **unbalanced**.

It contains only 4.8% fake job recruiting news,
and we should keep that in mind.

TABLE OF CONTENTS

01



Introduction

02



Preprocessing
&
EDA

03



NN architecture
&
Pre-trained Model

04



Other Findings
&
Conclusions

DATA-PREPROCESSING

Target

Fraudulent

Categorical Variables

Telecommuting
Has_company_logo
Has_questions
Employment_type
Required_experience
Required_education

Text Variables

Title
Department
Company_profile
Description
Requirements
Benefits
Industry
Function

Special Variables

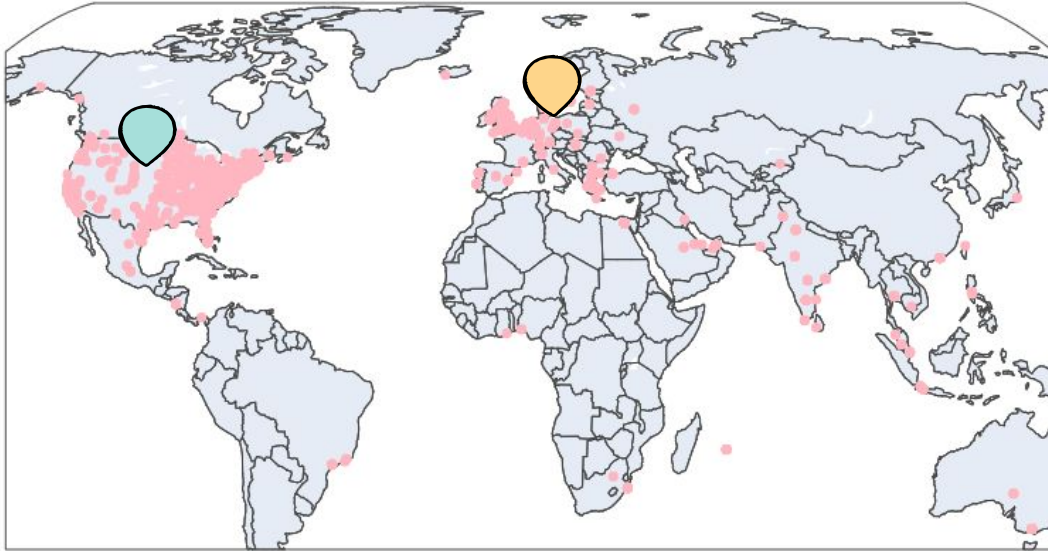
Salary_range
Location

600
tokens

17,879
postings

- 6 Categorical Variables - Give scales or get dummy
- 8 Text Variables - Combine and apply **TextVectorization()**
- 2 special variables
 - Salary range - Separate the range & get upper and lower bound
 - Location - Combine location with Text
- For the missing values, we fill in numerical ones as -1, text ones with 'unknown'

EDA - MAP OF LOCATION



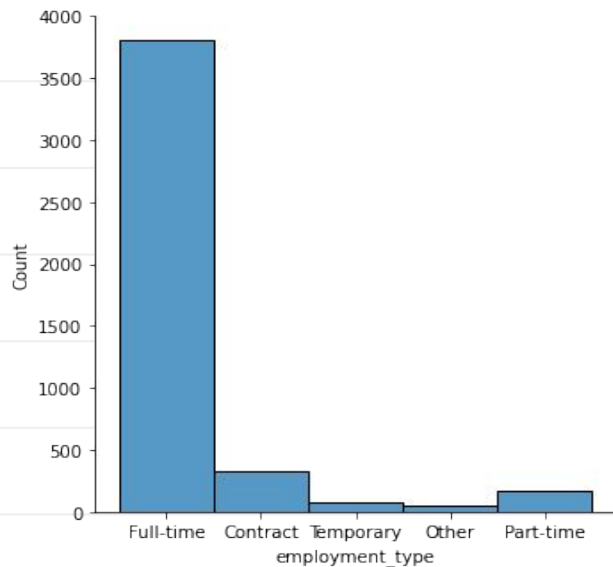
NORTH AMERICA



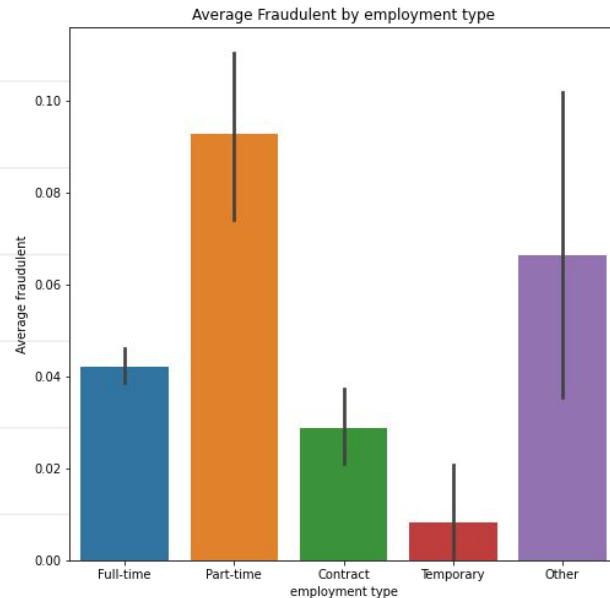
EUROPE

EDA - SOME MAJOR CATEGORICAL FEATURES

Employment type

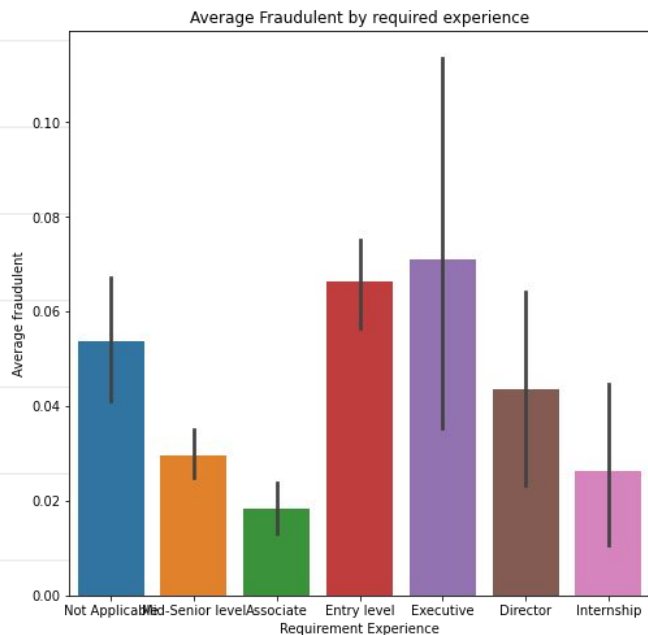


Avg Fraudulent by Employment Type



EDA - SOME MAJOR CATEGORICAL FEATURES

Avg Fraudulent by Required Experience



Avg Fraudulent by Requirement of Education

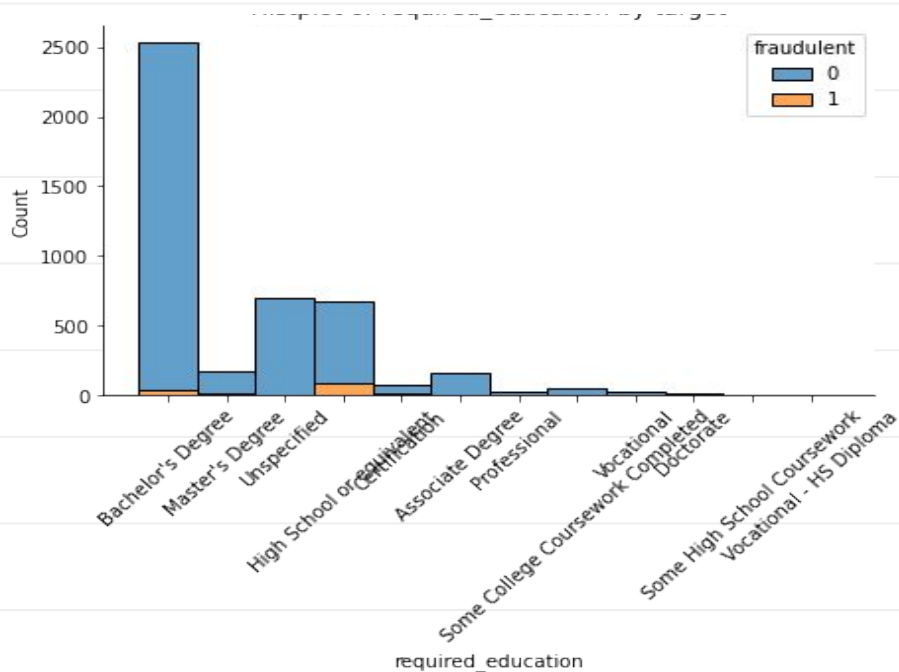


TABLE OF CONTENTS

01



Introduction

02



Preprocessing
&
EDA

03



NN architecture
&
Pre-trained Model

04



Other Findings
&
Conclusions

HOW WE PREPARE THE LAYERS FOR OUR MODEL

HOW WE DEAL WITH TEXT

- First, we used **TextVectorization() layer** to preprocess our text;
- Then we one-hot encoded these integer sequences but it crashed.

HOW WE DEAL WITH NUMERICAL

- After preprocessing, we added another layer to deal with all the numerical values.

NN ARCHITECTURE

MODEL SUMMARY

- Embedding layer & Numeric input
- Loss function - binary_crossentropy
- Train-Test Split - [:12516] (70/30%)
- Cross validation - k=4
- Activation - 'sigmoid', 'relu', 'linear'
- Optimizer - Rmsprop
- Dropout - 0.5
- Training-Accuracy: 97.91%
- Validation-Accuracy: 97.51%
- Test-set Accuracy - 94.89%

```
def build_model_embed():
    inputs = keras.layers.Input(shape=(1), dtype="string") # We take our strings as input
    processing = text_vectorization(inputs)

    # Truncates after 600 tokens, and pads up to 600 tokens for shorter reviews.
    # Mask zero means it will skip 0 tokens and will not pass them on.
    embedding = keras.layers.Embedding(input_dim=2000, output_dim=8, input_length=600, mask_zero=True)(processing)

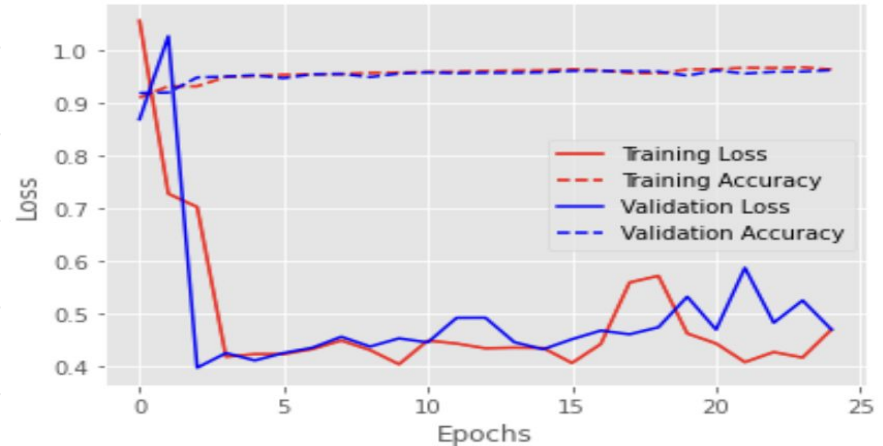
    x = keras.layers.Bidirectional(keras.layers.LSTM(32))(embedding)
    x = keras.layers.Dropout(0.5)(x)
    outputs = keras.layers.Dense(1, activation="sigmoid")(x)

    # Define numeric input branch
    input_numeric = keras.layers.Input(numeric_cate_subset.shape[1], name="Numbers")
    x = keras.layers.Dense(3, activation="relu")(input_numeric)
    x = keras.layers.Dense(3, activation="relu")(x)
    numeric_output = keras.layers.Dense(5, activation="linear")(x) #Another option might be to have dense matrices

    merge = keras.layers.Concatenate()([outputs, numeric_output])
    x = keras.layers.Dense(5, activation="relu")(merge)
    final_output = keras.layers.Dense(1)(x)

    model = keras.Model(inputs=[inputs, input_numeric], outputs = final_output)
    model.compile(optimizer="rmsprop", loss="binary_crossentropy", metrics=['accuracy'])
    return model

model = build_model_embed()
```



PRE-TRAINED EMBEDDINGS: GLOVE

MODEL SUMMARY

- Global Vector Representation
- Loss Function - binary_crossentropy
- Train-Test Split - [:12516]
- Cross validation - k=4
- Training-Accuracy: 95.09%
- Validation-Accuracy: 95.24%
- The resulting model is about 84.06% accurate in the holdout sample.

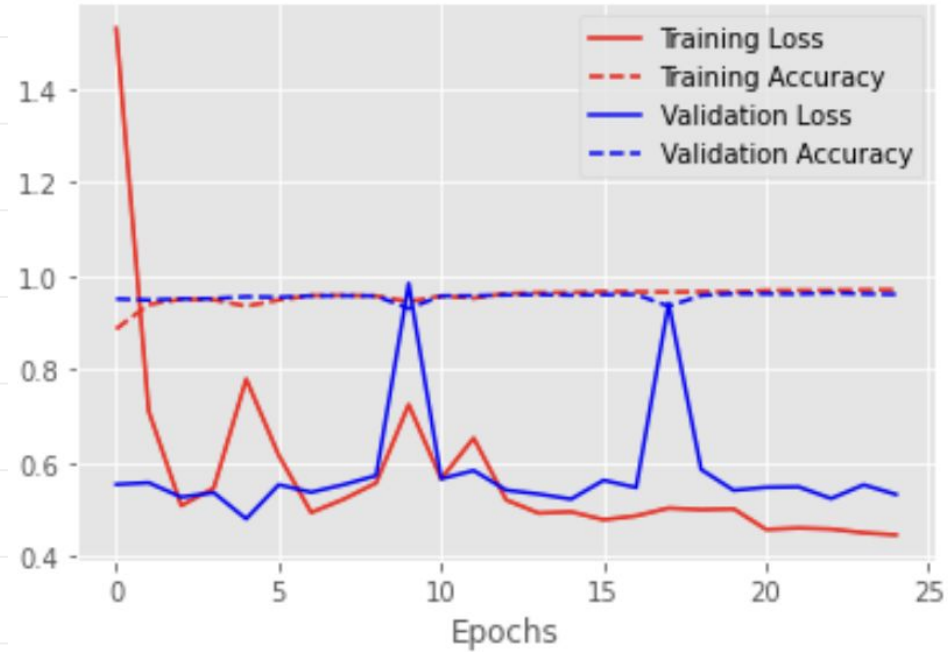


TABLE OF CONTENTS

01



Introduction

02



Preprocessing
&
EDA

03



NN architecture
&
Pre-trained Model

04



Other Findings
&
Conclusions

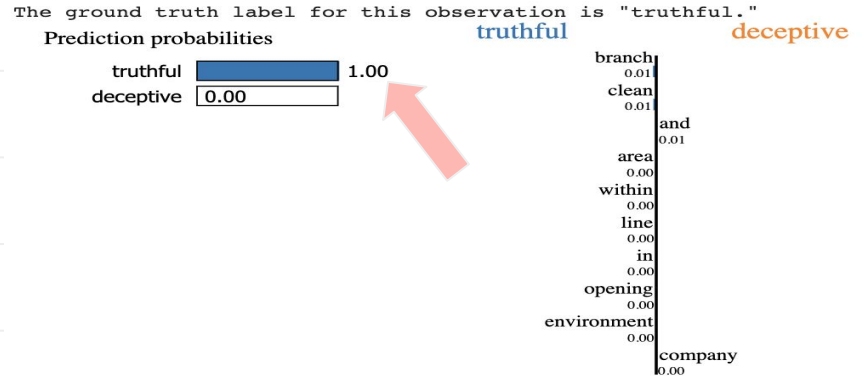
LIME

After doing a 20/80 test-train split, we trained our model processing text.

Accuracy = 98.34%

Randomly calling the one observation out, we tried multiple obs, we almost have 100% sure of predicting real job postings, around 99% probability of predicting fake ones.

Sample results showed as listed..

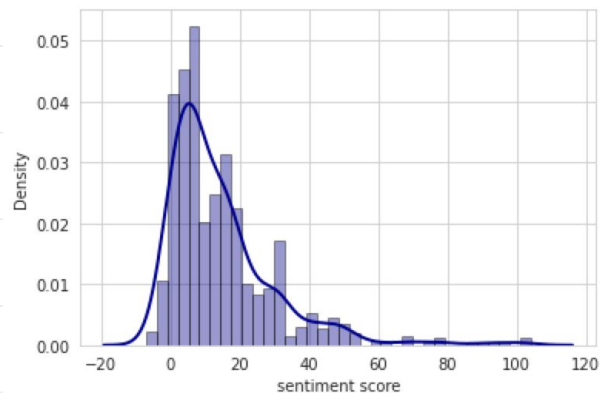


Text with highlighted words

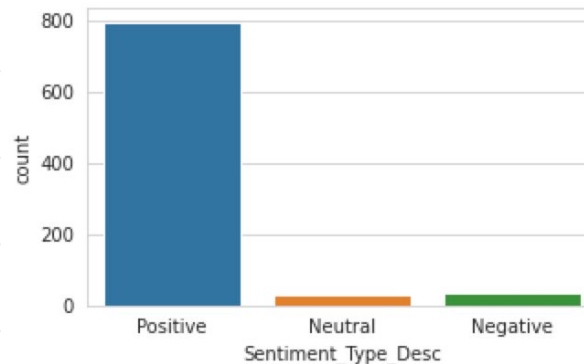
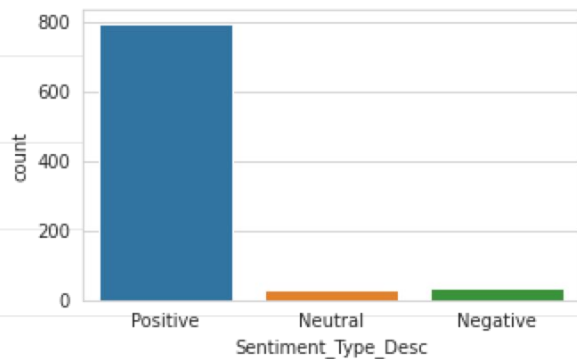
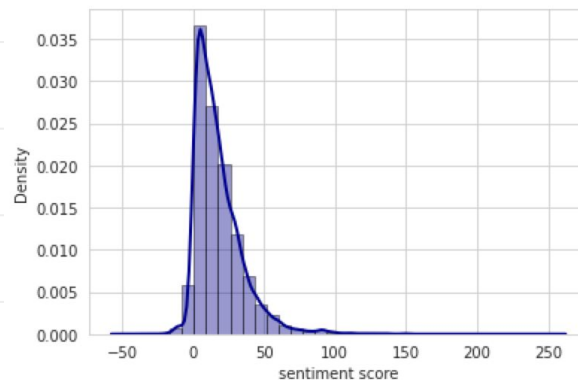
Medium Duty DriverUS, , All LocationsOur HistoryFounded in 1929 by Earl Bertrand Bradley, the company began selling products for Knap and Vogt Co. of Grand Rapids, Michigan. With the opening of the Los Angeles branch in 1929, the company became a wholesale distributor specializing in store fixture and specialty hardware such as drawer slides, hinges, brackets and standards. In 1943 branch offices were opened in San Francisco followed by Seattle in 1956. The company's market position and business began to grow during the late 1950's and 1960's after adding Wilsonart's high-pressure laminate line to its product mix. During the 1970's under the leadership of E.B. Bradley's son Robert E. Bradley, Sr., two new branches were opened; San Diego in 1972 and Portland in 1976. Significant product additions were the Blum line of European hinges and drawer slides and Accuride precision ball bearing drawer slides. A greater emphasis as a supplier to the Cabinet and Furniture industries was taking hold. Since the opening of the Anaheim branch in 1995, the company has been consistently growing. In 1998 the company entered the cold press lamination business by opening 3 locations of its West Coast Laminating subsidiary in the Los Angeles, San Francisco and Pacific Northwest marketplaces. Our OwnershipUp until January 4, 2009, the company operated as a 100% family owned business. On January 5, 2009, Industrial Opportunity Partners ("IOP"), a private equity firm based in Evanston, IL, partnered with Robert Bradlev, Jr. in acquiring the stock of E.B. Bradlev

SENTIMENT ANALYSIS

Fraud / Avg = 14.60



Non-Fraud / Avg = 17.98



WORD CLOUD

Company File



Benefits



It turns out that indeed our guesses are partially right. The fake ones do tend to promise bonus and benefits, supply with online training, guarantee work life balance.

WORD CLOUD

Job Description



Job Requirements



The fake job recruiting postings tend to describe their projects or products, we may assume they are just empty talk. For the requirements, they tend to stress experience - something can not be quantified.

CONCLUSION



MODEL 1

Our self designed model has **validation accuracy around 97%**.



MODEL 2

The pre-trained model has **validation accuracy around 95%**.



LIME

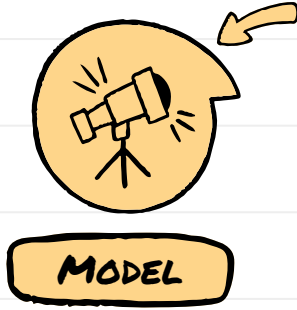
Only focused on text, the accuracy among test set is around 98%.



SENTIMENT
WORD CLOUD

Fake ones tend to talk big, guarantee great welfare.

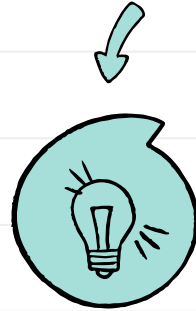
SUGGESTIONS



We tried to implement one-hot encoding processing the integer sequences, only it got crashed.

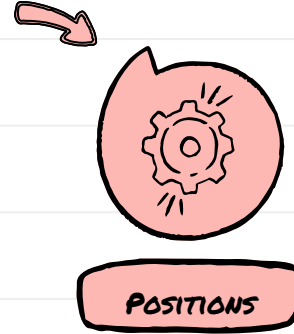
But for the first glance, the performance after one-hot is better.

For the embeddings, we can increase the depth of it (utilize more tokens).



Fake job recruiting postings tend to focus on entry level or executive level; part-time job.

Be cautious when applying for these types of positions.

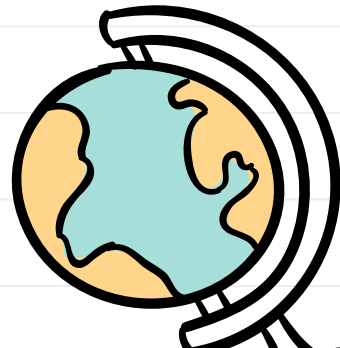
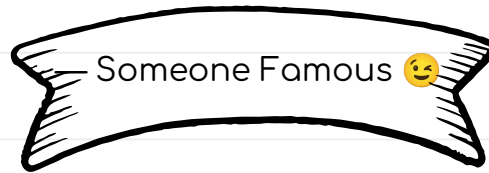


After doing sentiment analysis and word cloud towards fake ones, we figured out that they tend to talk big about themselves, but avoid some essential information.

When seeing great welfare and work environment, be cautious.



**"THINK POSITIVE.
BE CAUTIOUS."**



**THANKS
FOR
LISTENING!**

- Team REAL -

