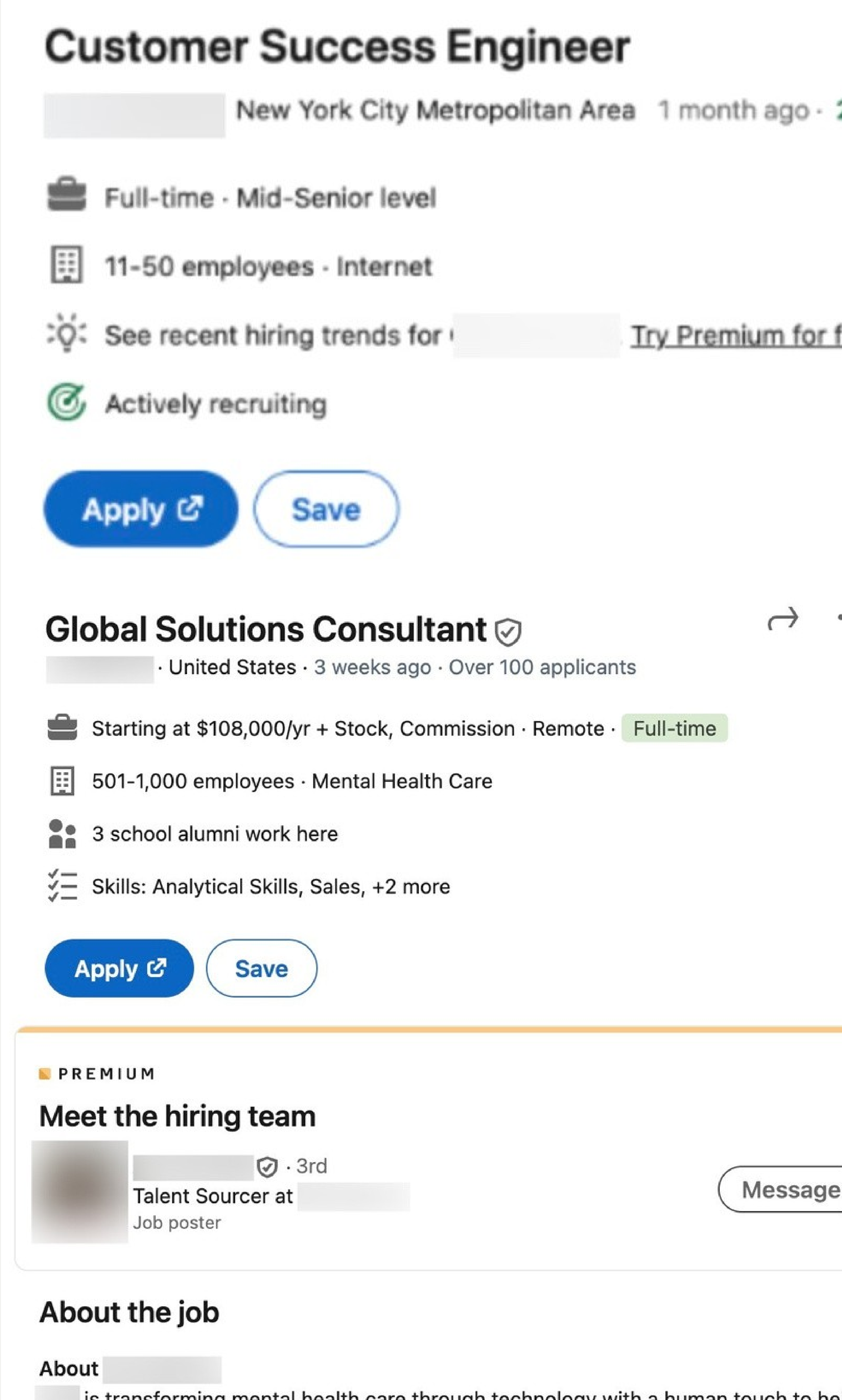
Job Listing “Integrity“ Investigation

By: Dian Jin, Mingze Wu, Tanvi Sheth, Sneha Ekka, Jenil Shah



Customer Success Engineer

[redacted] New York City Metropolitan Area 1 month ago ·

Full-time · Mid-Senior level

11-50 employees · Internet

See recent hiring trends for [redacted] Try Premium for f

Actively recruiting

Global Solutions Consultant

[redacted] · United States · 3 weeks ago · Over 100 applicants

Starting at $108,000/yr + Stock, Commission · Remote · Full-time

501-1,000 employees · Mental Health Care

3 school alumni work here

Skills: Analytical Skills, Sales, +2 more

PREMIUM

Meet the hiring team

[redacted] · 3rd

Talent Sourcer at [redacted]

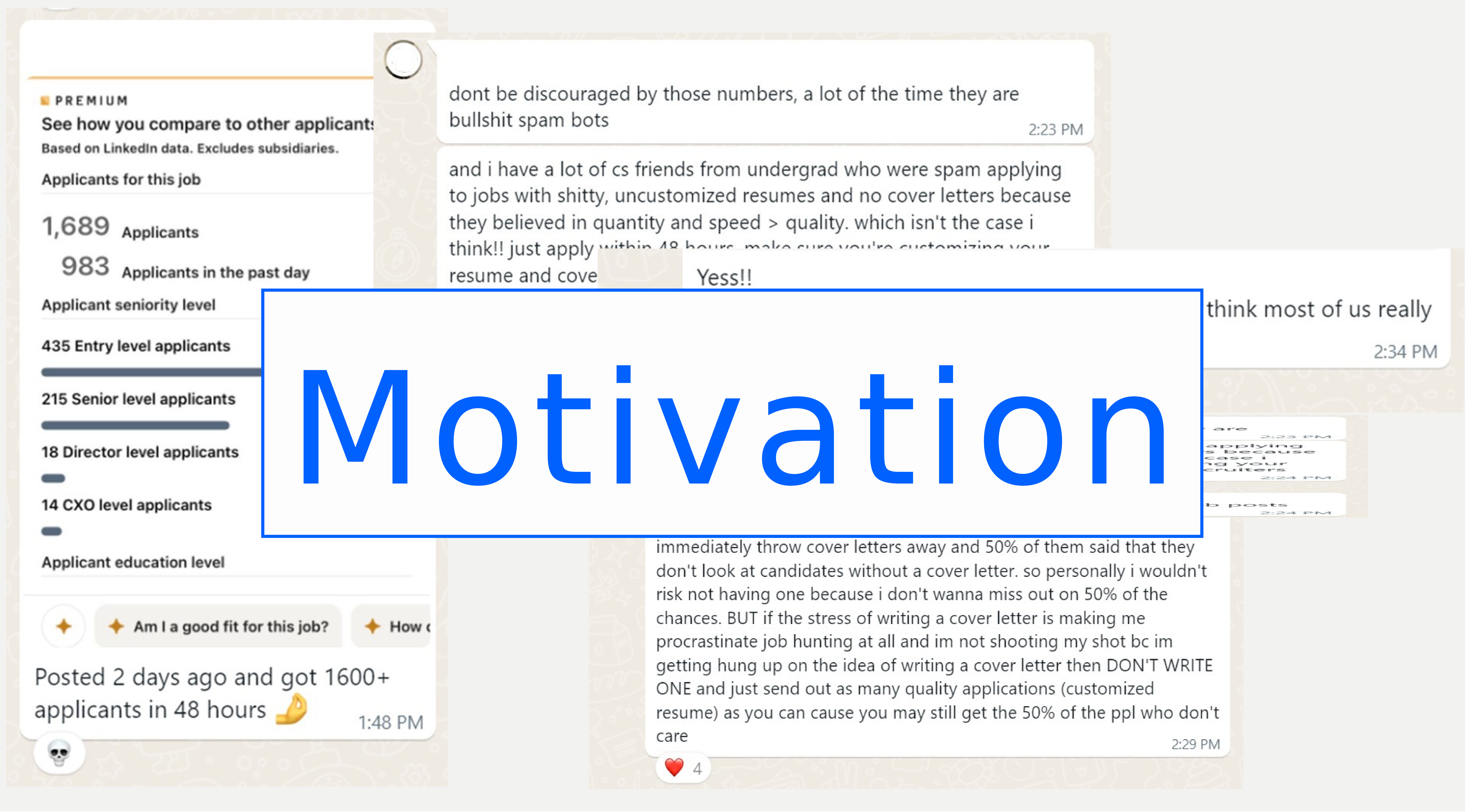
Job poster

About the job

About [redacted]

[redacted] is transforming mental health care through technology with a human touch to he

Motivation



PREMIUM

See how you compare to other applicants

Based on LinkedIn data. Excludes subsidiaries.

Applicants for this job

1,689 Applicants

983 Applicants in the past day

Applicant seniority level

435 Entry level applicants

215 Senior level applicants

18 Director level applicants

14 CXO level applicants

Applicant education level

Posted 2 days ago and got 1600+ applicants in 48 hours 🤌 1:48 PM

💀

dont be discouraged by those numbers, a lot of the time they are bullshit spam bots 2:23 PM

and i have a lot of cs friends from undergrad who were spam applying to jobs with shitty, uncustomized resumes and no cover letters because they believed in quantity and speed > quality. which isn't the case i think!! just apply within 48 hours, make sure you're customizing your resume and cove

Yess!!

think most of us really

2:34 PM

are 2:23 PM

applying s because case i ng your cruiters 2:24 PM

b posts 2:24 PM

immediately throw cover letters away and 50% of them said that they don't look at candidates without a cover letter. so personally i wouldn't risk not having one because i don't wanna miss out on 50% of the chances. BUT if the stress of writing a cover letter is making me procrastinate job hunting at all and im not shooting my shot bc im getting hung up on the idea of writing a cover letter then DON'T WRITE ONE and just send out as many quality applications (customized resume) as you can cause you may still get the 50% of the ppl who don't care 2:29 PM

❤️ 4

INTRODUCTION

Motivation

01

Authenticity of job postings MATTERS

02

Spam/fake posts everywhere

03

Important to build a trustworthy environment for recruiters and candidates

INTRODUCTION

Problem Statement

01

Process textual job posting information and draw out patterns that distinguish fraudulent jobs from real ones

02

Offer an automated and reliable solution that can enhance the detection of fraudulent postings

INTRODUCTION

Data Overview

01

Real / Fake Job Posting
Prediction Dataset from Kaggle

02

Collected by The University of the Aegean | Laboratory of Information & Communication Systems Security

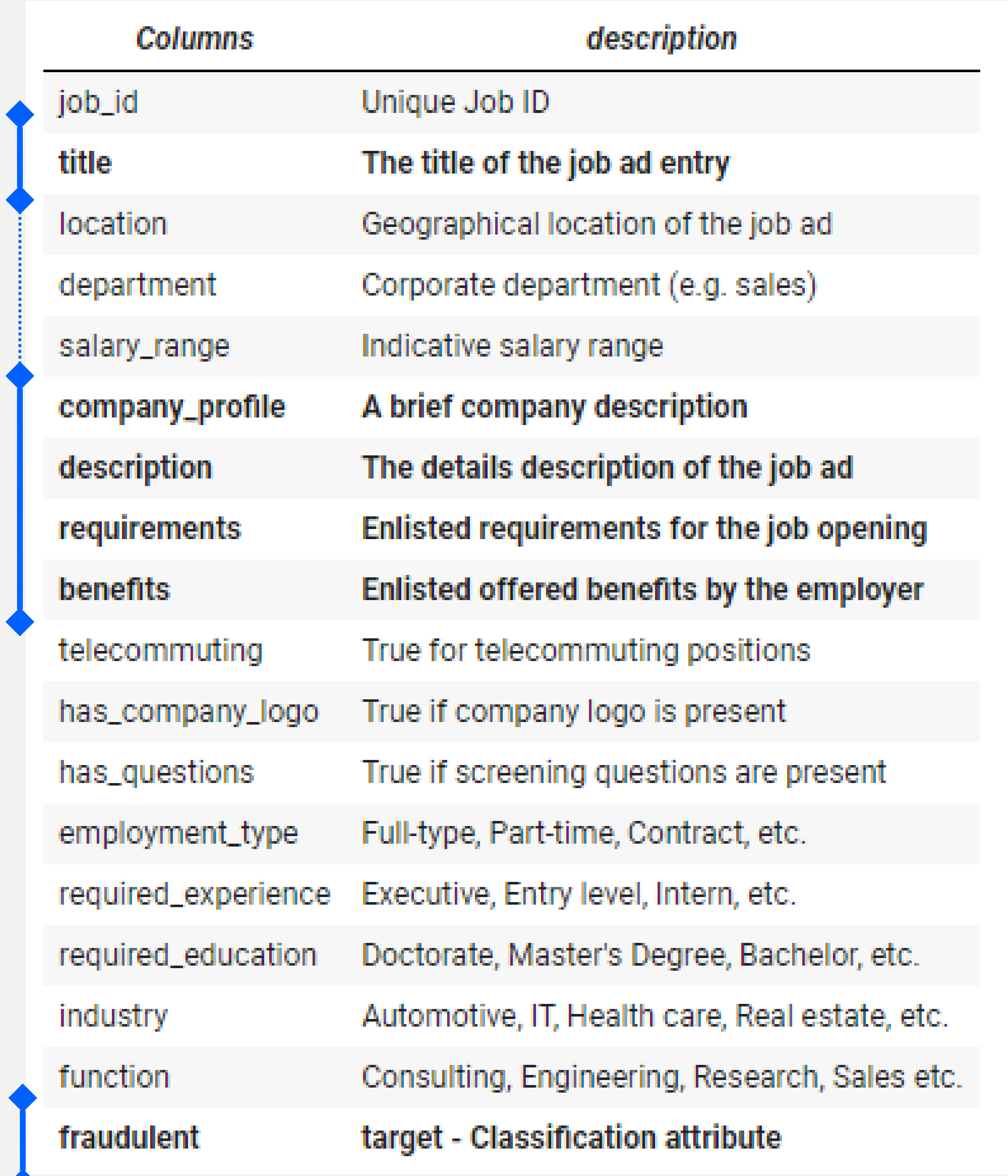
03

17880 job postings

04

18 features

KEY FEATURES



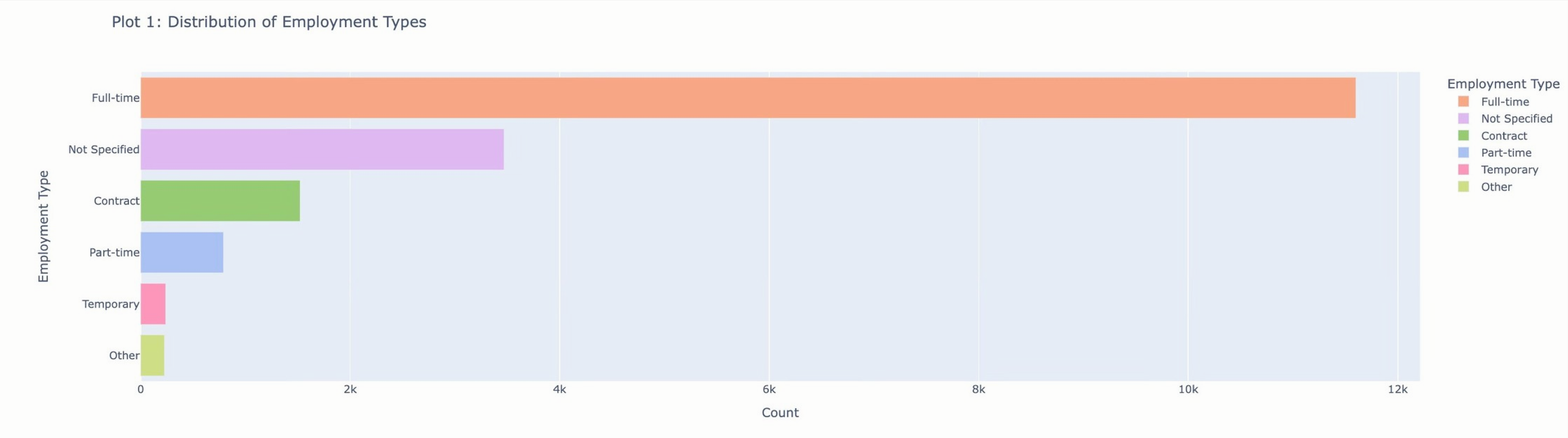
Columns	*description*
job_id	Unique Job ID
title	**The title of the job ad entry**
location	Geographical location of the job ad
department	Corporate department (e.g. sales)
salary_range	Indicative salary range
company_profile	**A brief company description**
description	**The details description of the job ad**
requirements	**Enlisted requirements for the job opening**
benefits	**Enlisted offered benefits by the employer**
telecommuting	True for telecommuting positions
has_company_logo	True if company logo is present
has_questions	True if screening questions are present
employment_type	Full-type, Part-time, Contract, etc.
required_experience	Executive, Entry level, Intern, etc.
required_education	Doctorate, Master's Degree, Bachelor, etc.
industry	Automotive, IT, Health care, Real estate, etc.
function	Consulting, Engineering, Research, Sales etc.
fraudulent	**target - Classification attribute**

TEXTUAL INFORMATION

TARGET

GRAPHS

Data Insights



Majority Full-time employees

Over 11,000 postings offer full-time positions

Unspecified employment types

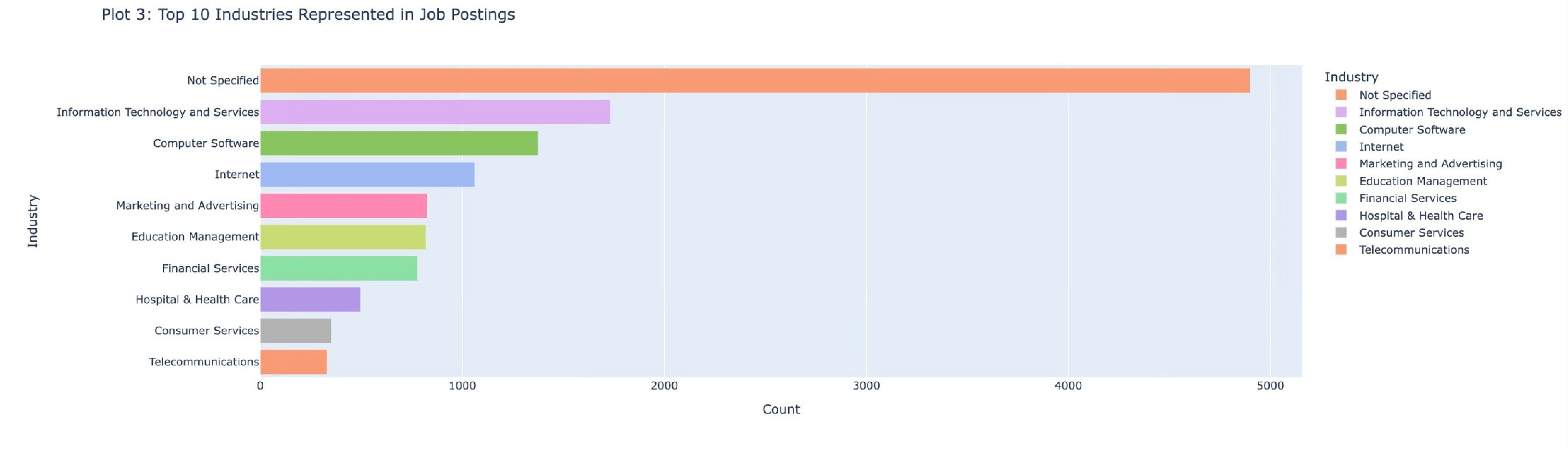
Significant portion of the workforce

16.8% of positions are non-full time roles

Contract, part-time, temporary, and other roles constitute about 3,000 positions

GRAPHS

Data Insights



Majority of ‘Not Specified’ job postings

Close to 5000 postings show ‘Not Specified’ job positions

Technology represents about 23.35% jobs

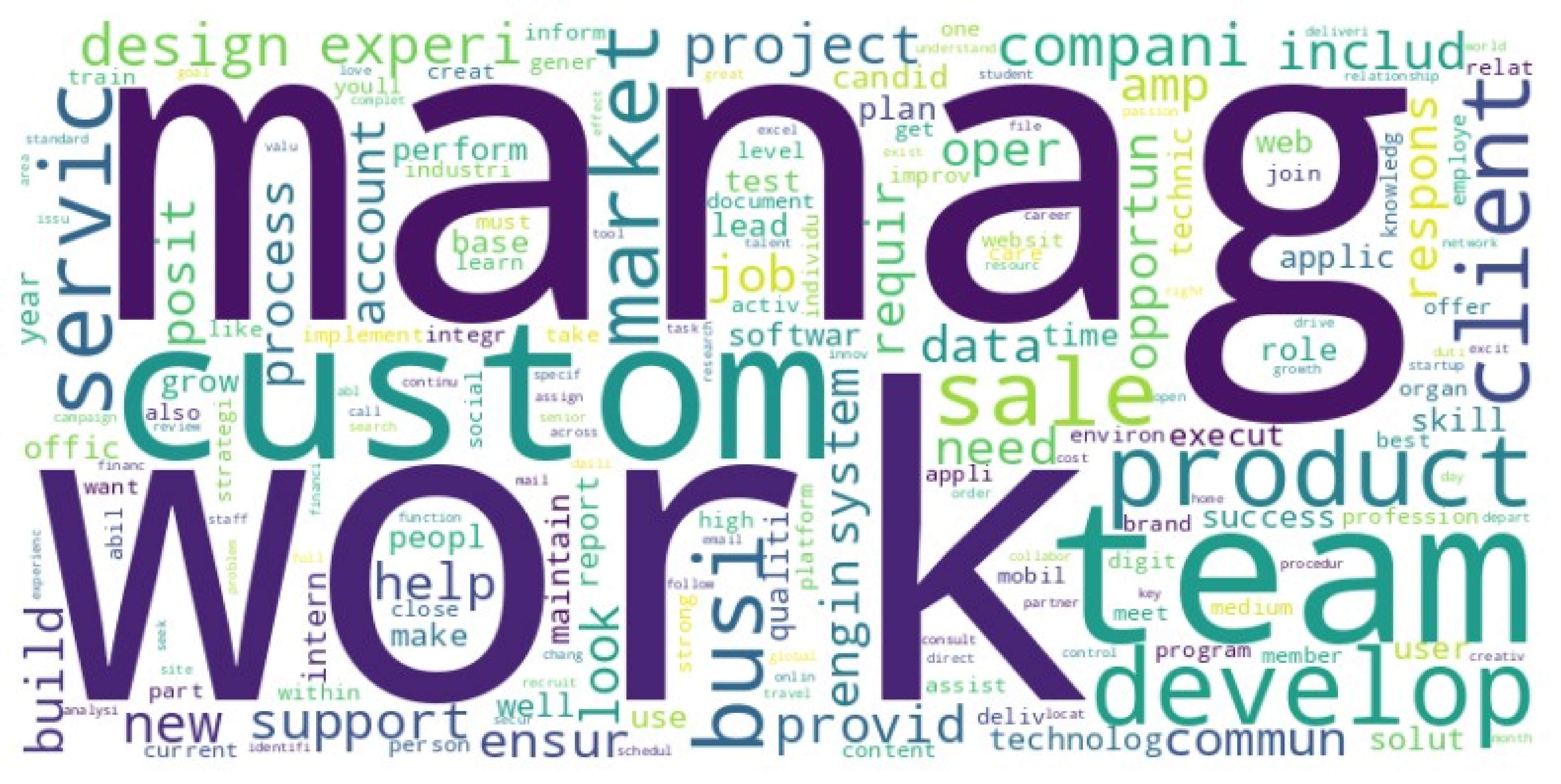
Infomation Technology and Services, Computer Software and Internet represents top 3 industries

Wide ranges of industries hiring!

About 9 industries are actively hiring.

WORD CLOUD

Data Insights

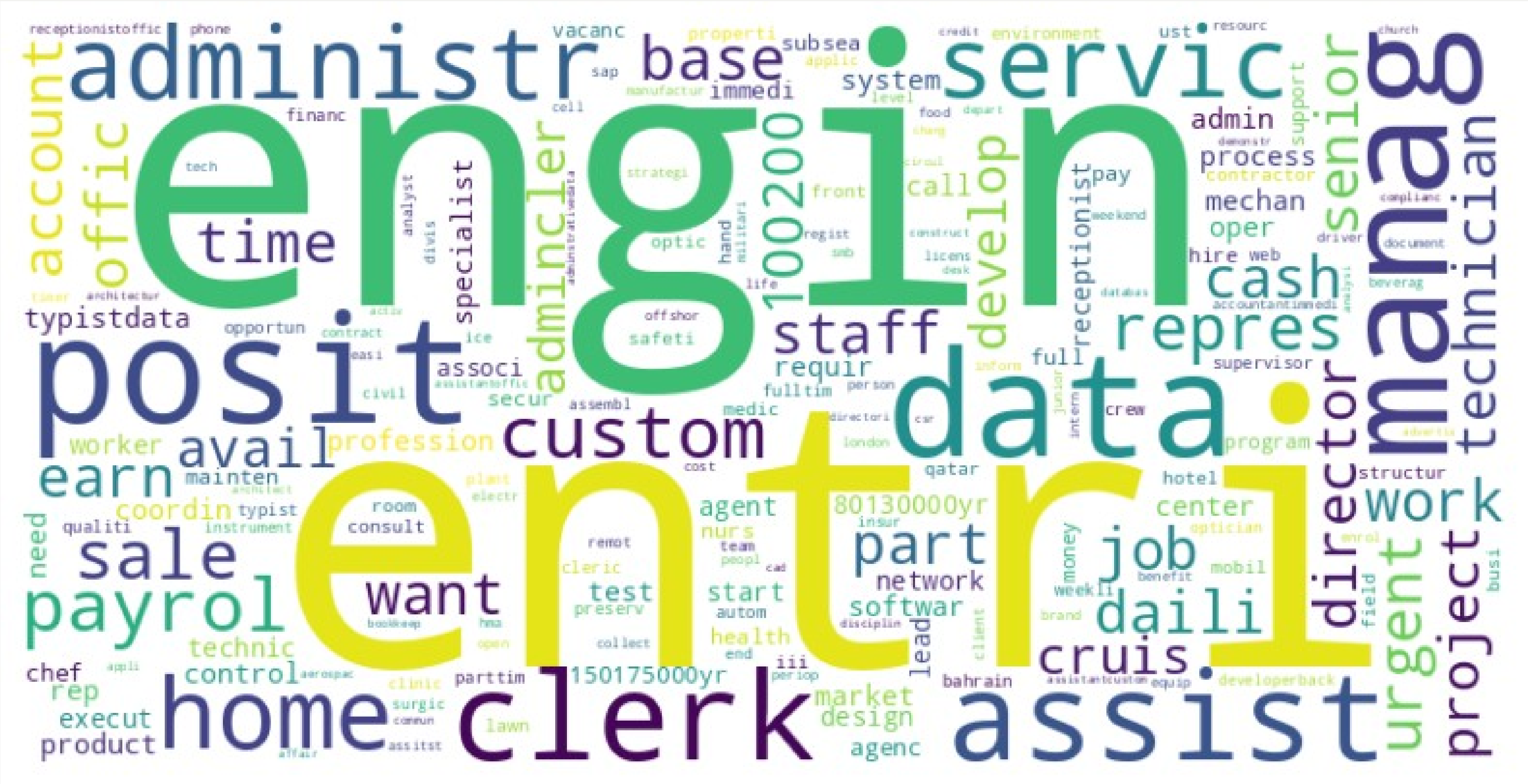


Popular Job Descriptions

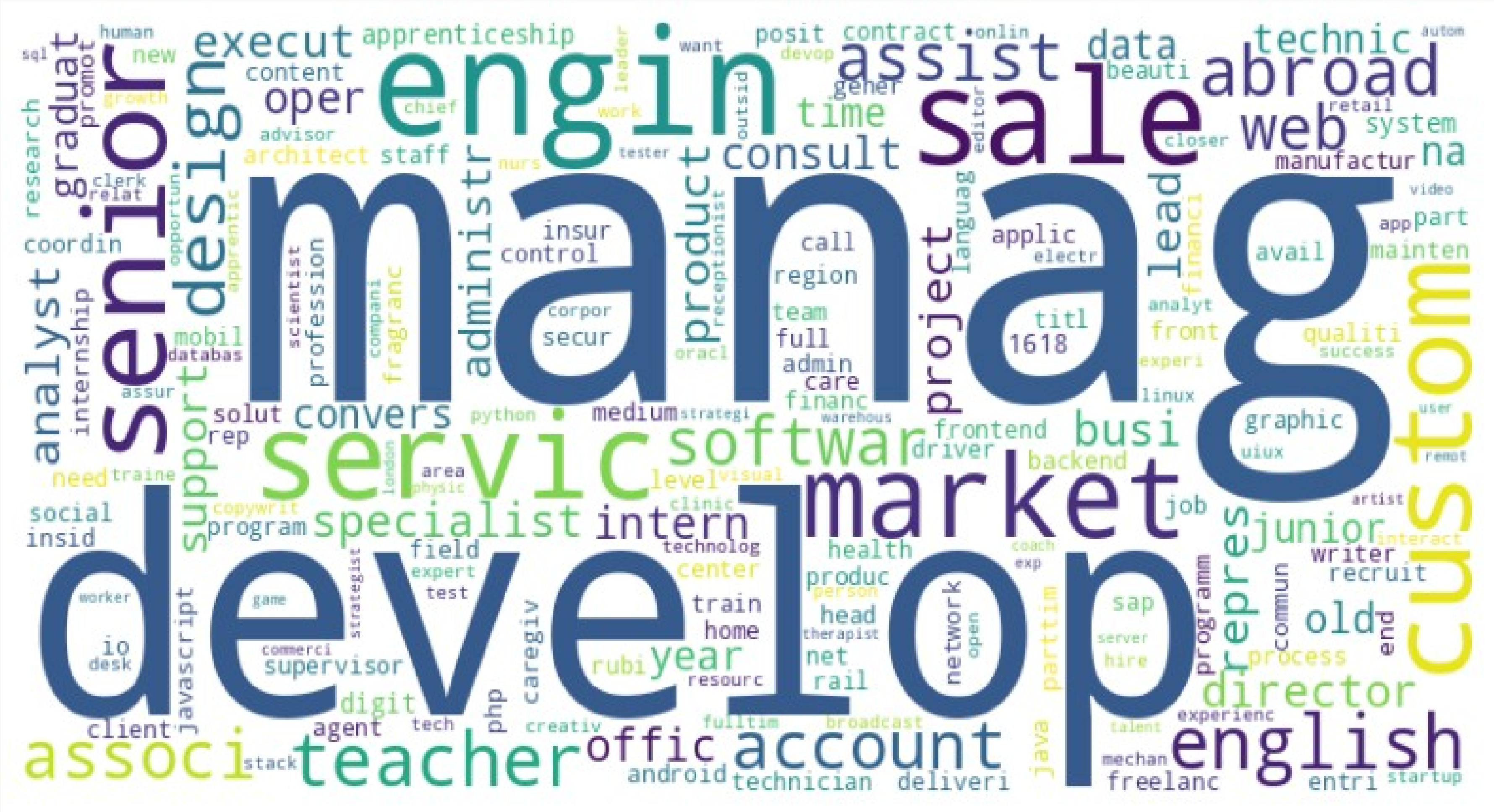
Popular Job Requirements

WORD CLOUD

Fake vs Real Jobs



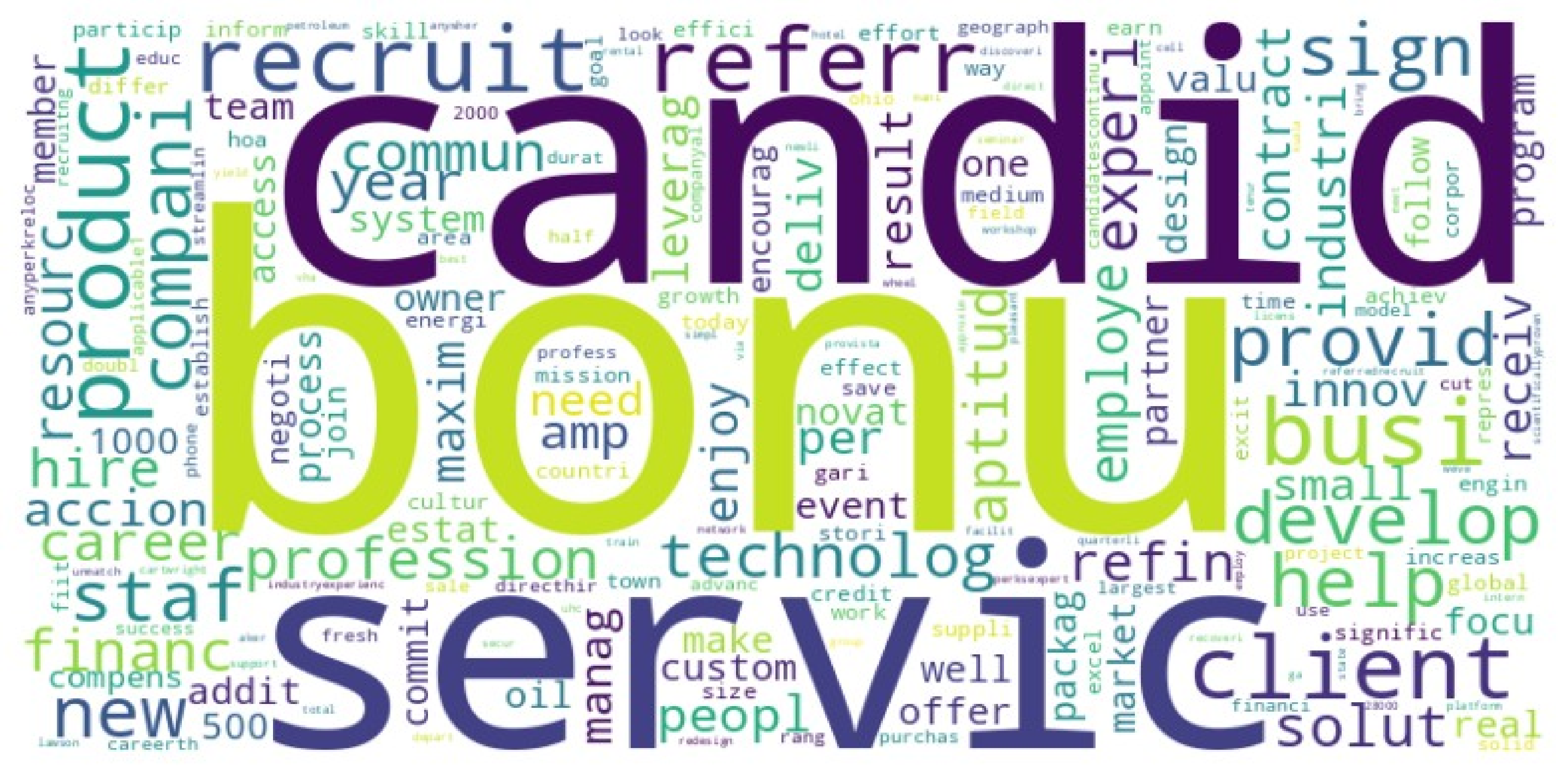
“Fake” Titles



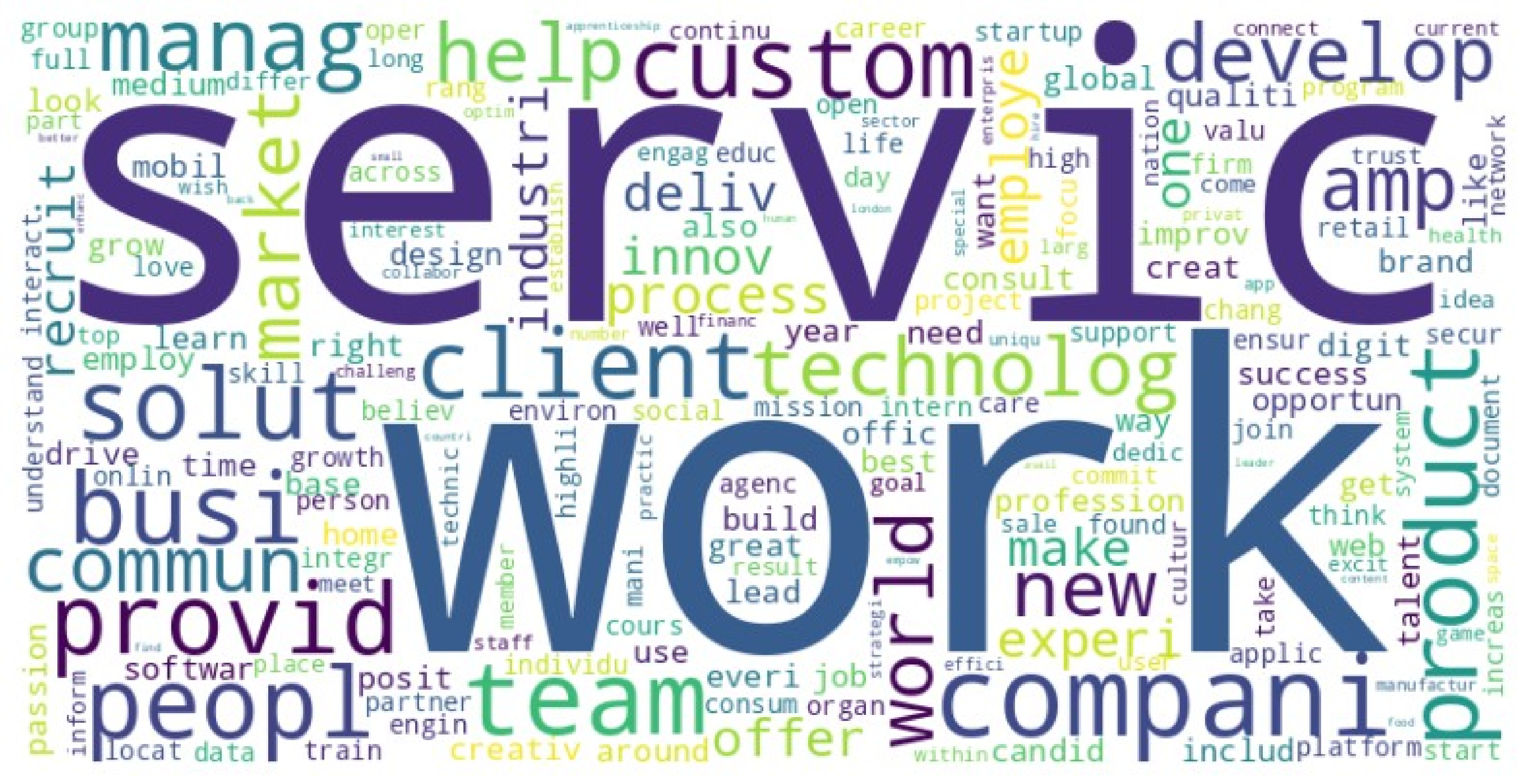
“Real” Titles

WORD CLOUD

Fake vs Real Jobs



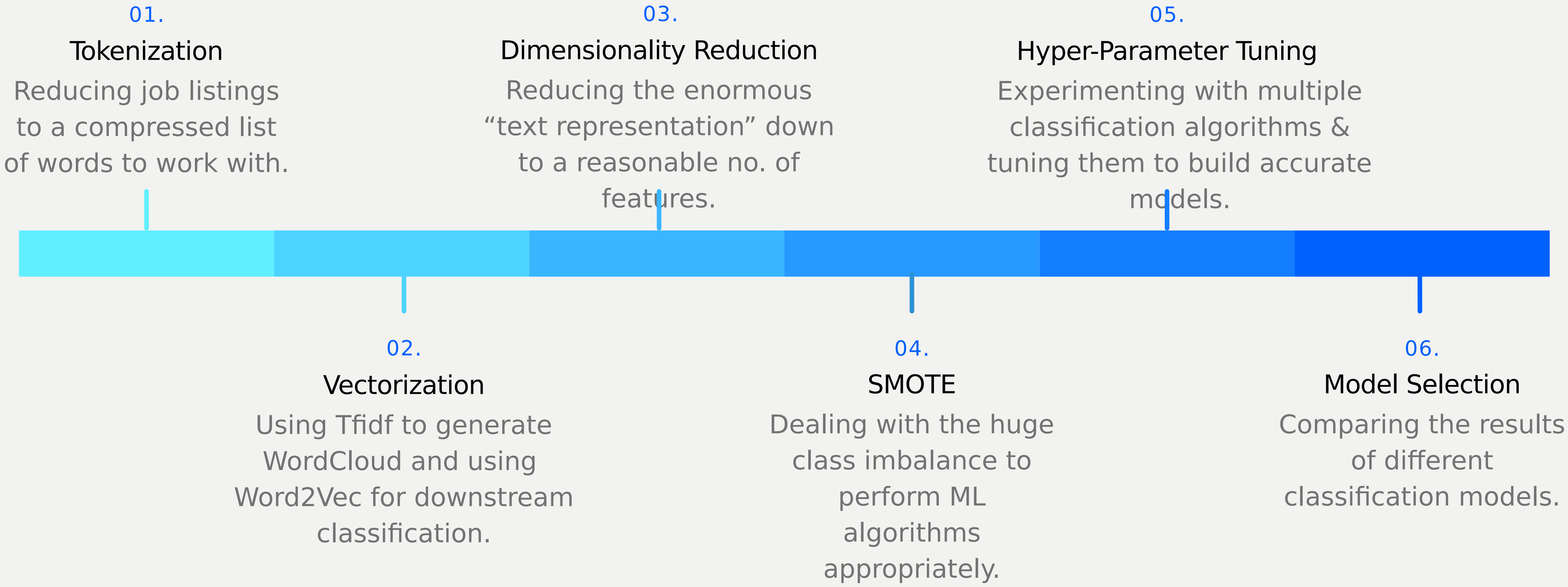
“Fake” Company Profiles



“Real” Company Profiles

ANALYSIS & MODELING

Project Plan



01.

Tokenization

Reducing job listings to a compressed list of words to work with.

02.

Vectorization

Using Tfidf to generate WordCloud and using Word2Vec for downstream classification.

03.

Dimensionality Reduction

Reducing the enormous “text representation” down to a reasonable no. of features.

04.

SMOTE

Dealing with the huge class imbalance to perform ML algorithms appropriately.

05.

Hyper-Parameter Tuning

Experimenting with multiple classification algorithms & tuning them to build accurate models.

06.

Model Selection

Comparing the results of different classification models.

ANALYSIS & MODELING

Comparing Results

	Logistic Regression	Random Forest	SVC	XGBoost	Stacking
F1 Score	0.468	0.764	**0.816**	0.803	0.762
Balanced Accuracy	0.884	0.860	0.864	**0.874**	0.809

CONCLUSION

Challenges

“salary_range”

Missing and invalid values

Multiple currencies

Varying country standards

Clustering

Assumed clustering would help identify fraudulent job listings

CONCLUSION

Limitations

‘word2vec’ Model

Hyper-parameter selection

Limited time and resources

Used pre-defined combinations for arguments

Class Imbalance

4.8% of data is fraudulent

Implemented SMOTE

Unable to verify effectiveness

Possibility of Overfitting

Vectorizer Comparisons

word2vec VS GloVe

Inability to find pre-trained models on job listings

CONCLUSION

Implications

Verifying Legitimacy of Job postings

Using this model, platforms can verify the legitimacy of job postings to maintain the trust of users

Starting point for a more complex model

Can build on this to create a more intricate model with hyper-parameters and train it on data from all different job posting platforms

Thank You!

Questions?