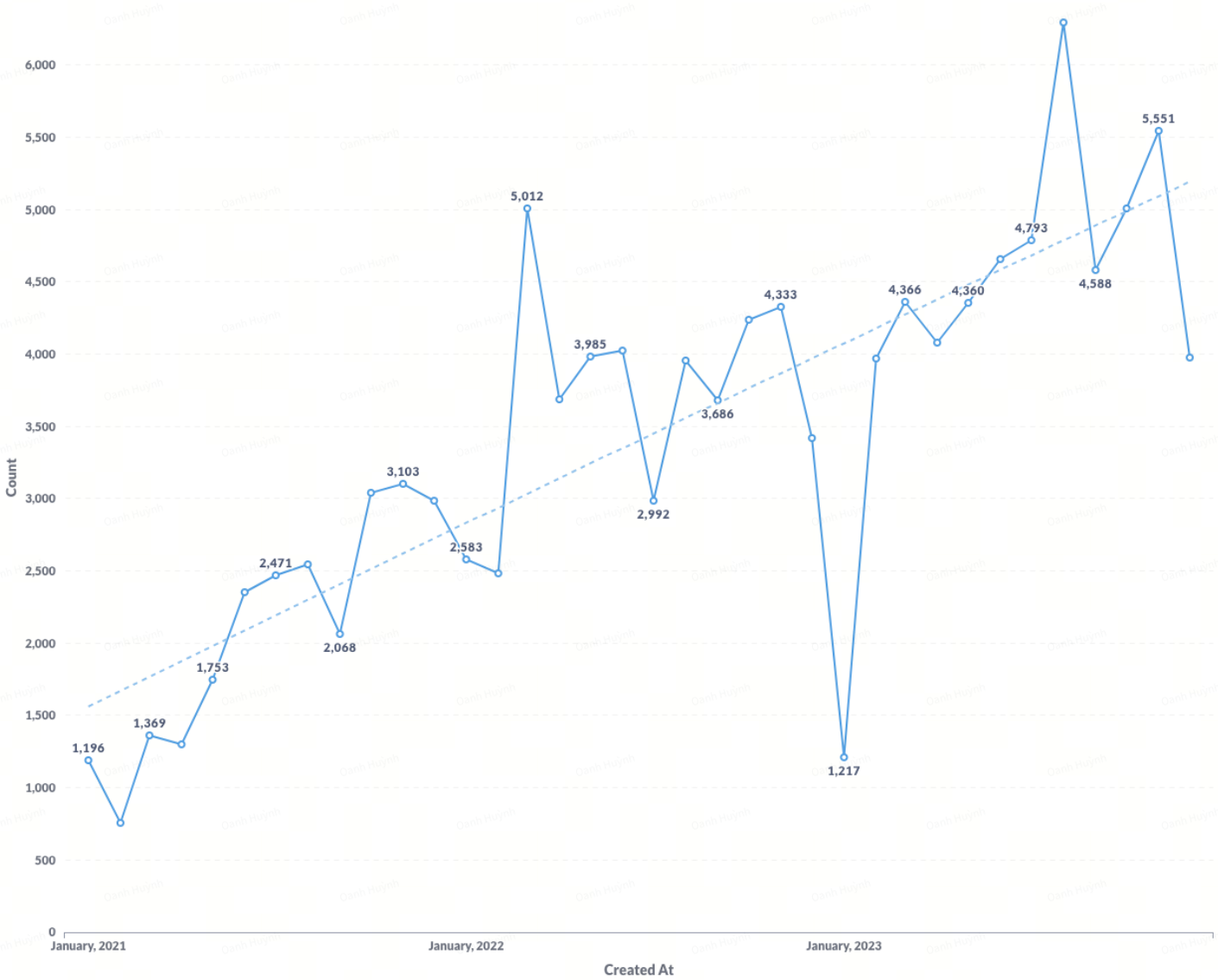
Duplicated Job Check Tool

1. The context

Since we have lacked control over job quality since 2021 due to limited manpower, a lot of complaints about job quality arose, along with the new job created numbers increasing on our job platform.



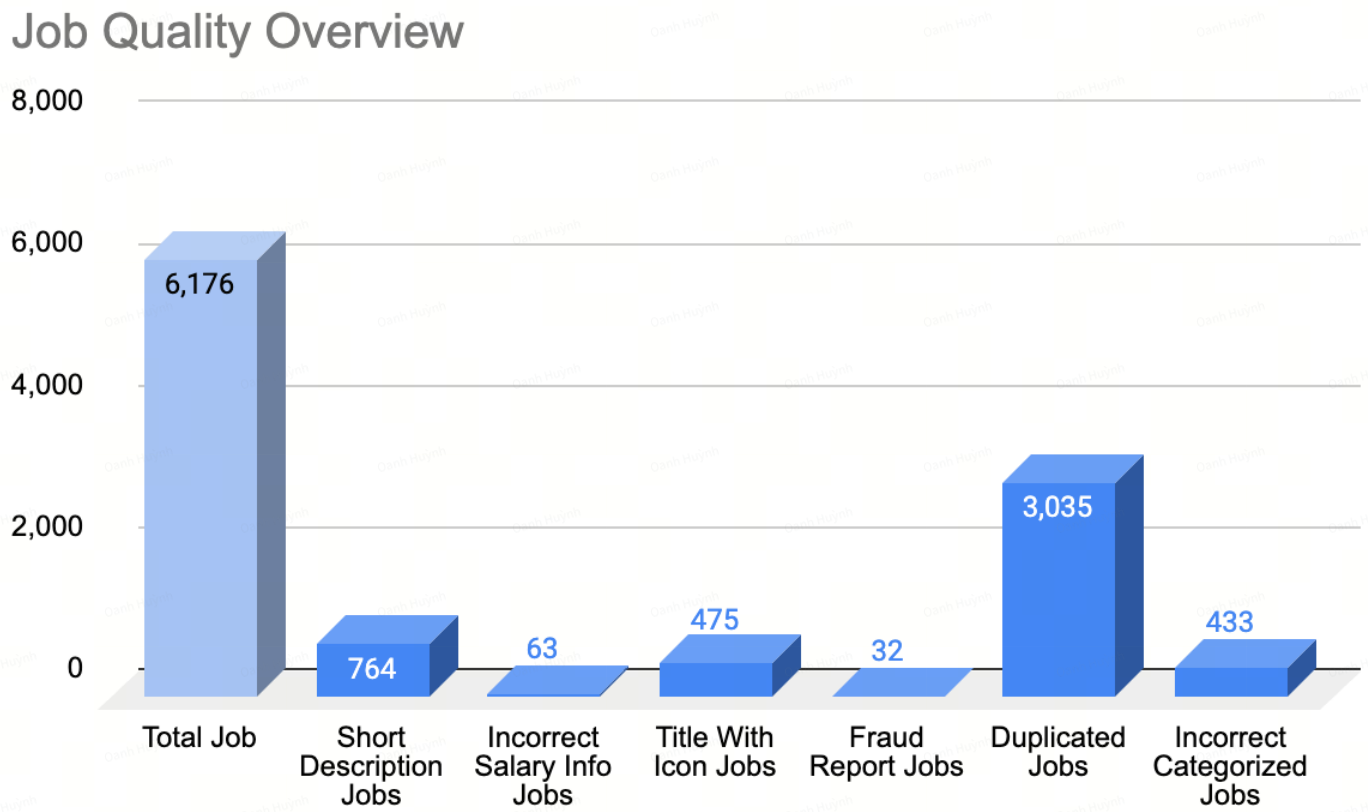
Monthly job created from Jan 2021 to Dec 2023

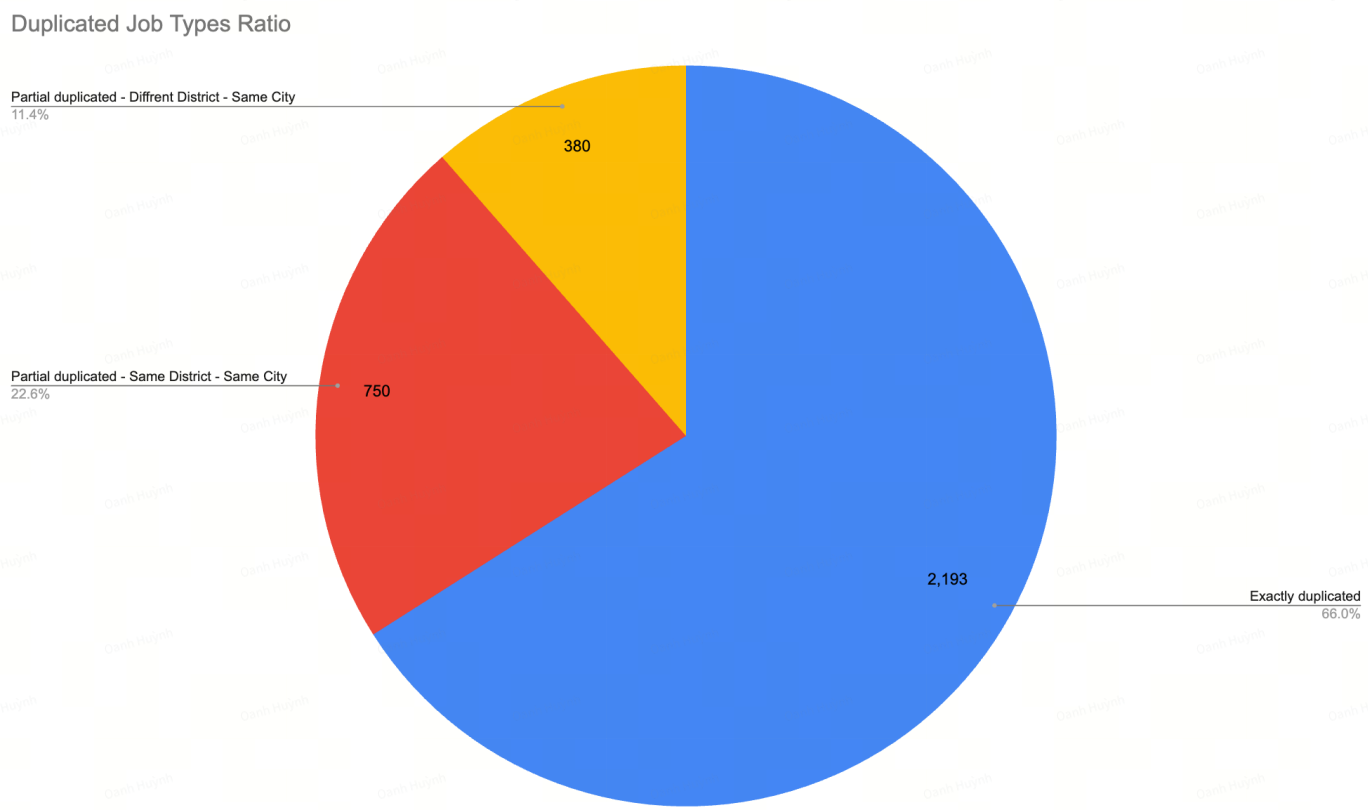
2. The first analysis of Job Quality

2.1 Data source

- Date range: Jobs created last 30 days from 6 Dec 2023 in VN
- Created a criteria checklist for Job Quality from user complaints and internal discussion, here're a few criteria for the analysis:
  - Job
    - Duplicated Jobs
    - Fraud Report Jobs
    - Incorrect Categorized Jobs
    - Incorrect Salary Info Jobs
    - Short Description Jobs
    - Title With Icon Jobs
  - Company
    - Rejected / Unverified Companies
    - No Tax Number Companies
    - No Logo Companies

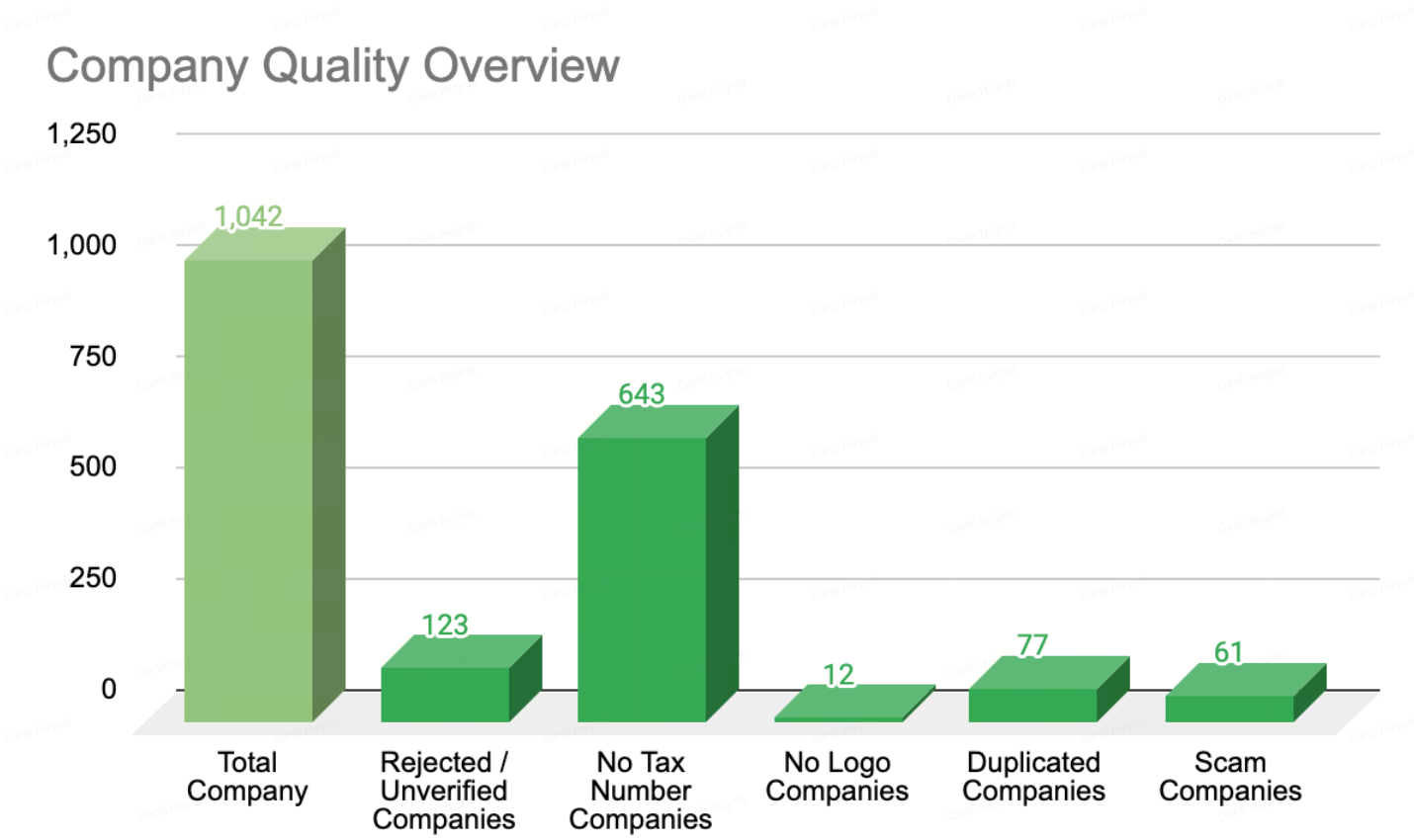
2.2 Analyze data





There're 3 types of duplicated jobs here:

1. Exactly duplicated:
    a. Exactly same: Job Title, Job Description, Job's Company Name, Job's Location
2. Partial duplicated - Same District - Same City:
    a. Similar: Job Title, Job Description
    b. Exactly same: Job's Company Name, Job's Location - District level
3. Partial duplicated - Different District - Same City:
    a. Similar: Job Title, Job Description
    b. Exactly same: Job's Company Name, Job's Location - City level



2.3 Actionables

3. Main focus - Duplicated Job Check

3.1 Ver 1 - Manually check

3.2 Ver 2 - Python Script

Included:

1. Exactly Duplicated Jobs Check
    a. Exact job_title
    b. Exact job_description
    c. Exact job_district
    d. Exact company_name
2. Partial Duplicated - Same District Jobs Check
    a. Exact job_category
    b. Exact job_district
    c. Exact company_name
3. Partial Duplicated - Different District Jobs Check
    a. Exact job_category

    b. Different job_district
    c. Exact job_city
    d. Exact company_name
4. Duplicated Companies Check
    a. Exact company_name
    b. Exact company_city

3.3 Ver 3 - Python Script - Optimized

Since the checker is still overcounted compared with the actual number of duplicated jobs, we decided to improve the checker to be more accurate.

Reason

Investigated on the incorrect cases by the checker, we found out some reasons behind the inaccurate:

1. No condition for Job Type in the checker:
    a. There are same-title jobs but in different types: FULL_TIME, PART_TIME and PROJECT_BASED, which are actually not considered a duplicated case.
2. Jobs in the same category, but actually are different job titles.
    a. Example:
        i. Job Title 1 : "Accountant - Internal" in Category "Accountant"
        ii. Job Title 2 : "Accountant - Outsourcing" in Category "Accountant"
    b. In this example, both are in "Accountant" category, but the working responsibilities and team are different, which having different Job Description

Adjustment

Based on these 2 reasons, we modified the script with:

1. Added different job_type (full-time, part-time, intern) condition
2. Added 80% Similarity in Job Description condition

Result

We did the manual check to have the Actual Duplicated Jobs numbers for 4 weeks. Result of checker optimization:

- Before change:
    - ***Checker*** *- % Duplicated Jobs/ Total Open Jobs:* 40.39%

- ***Actual*** *- % Duplicated Jobs/ Total Open Jobs:* 23.62%

- After change:
  - ***Checker*** *- % Duplicated Jobs/ Total Open Jobs:* 16.70%
  - ***Actual*** *- % Duplicated Jobs/ Total Open Jobs:* 16.45%

Metrics	Last 10W Trendline	% WoW Change	% MoM Change	% QoQ Change	% YoY Change	22 Apr - 28 Apr	29 Apr - 5 May	6 May - 12 May	13 May - 19 May	20 May - 26 May
% Duplicated Jobs/Open Jobs (est)		-59%	-60%	-61%	-	43.75%	39.92%	42.48%	40.39%	16.70%
% Duplicated Jobs/Open Jobs (act)		-30%	-32%	-32%	-	-	23.53%	25.22%	23.62%	16.45%

- *% Duplicated Jobs/Open Jobs (est): from the checker*
- *% Duplicated Jobs/Open Jobs (act): actual manual check*

Similarity 80%

Pros

- Can quick spot on these cases:
  - Shuffle title but the job description is similar
  - Title EN - VN but the job description is similar
- It's a better bypass job with a similar title but actually different jobs - Since the descriptions are different. Example:
  - Nhân viên kinh doanh kênh otc khu vực cần thơ
  - Nhân viên kinh doanh etc khu vực cần thơ

Cons

- Can't spot on cases:
  - Short description
  - Mix description (Same description but different arrangement)
  - Shuffled title and description
  - 1 job - split into 2 jobs, each job has a half-description
- Low spot rate on fraud cases:
  - Since fraud would have shuffled title and description jobs, it didn't catch the fraud very well.
  - Can build a machine learning model to detect fraud cases

Accuracy check

We will use this accuracy model below to check on 2 versions:

1. Version 1:

a. Exactly same: category_l3, company_name, job_district

2. Version 2:
   a. Exactly same: category_l3, company_name, job_district, job_type
   b. 80% similar in: job_description

The Acuraccy Metrics		
	Predicted as Positive	**Predicted as Negative**
Actual: Positive	True Positive (TP)	False Negative (FN)
Actual: Negative	False Positive (FP)	True Negative (TN)

	Predicted as Positive	**Predicted as Negative**
Actual: Positive	True Positive Rate (TPR):	False Negative Rate (FNR) Called Miss Detection Rate (tỉ lệ bỏ sót)
Actual: Negative	False Positive Rate (FPR) Called as False Alarm Rate (tỉ lệ báo động nhầm)	True Negative Rate (TNR)

1st Check - Date: 05/28/24

Ver 1 - Both Scam and Duplicated		
	Predicted as Positive	**Predicted as Negative**
Actual: Positive	1,496	13
Actual: Negative	1,750	4,8

	Predicted as Positive	**Predicted as Negative**
Actual: Positive	91.61%	8.39
Actual: Negative	26.45%	73.5

Ver 2 - Similarity 80 - Both Scam and Duplicated		
	Predicted as Positive	**Predicted as Negative**
Actual: Positive	1,076	55
Actual: Negative	312	6,3

	Predicted as Positive	**Predicted as Negative**
Actual: Positive	65.89%	34.1
Actual: Negative	4.72%	95.2

2nd Check - Date: 06/10/24

Ver 1 - Both Scam and Duplicated		
	Predicted as Positive	Predicted as Negativ
Actual: Positive	1,361	2
Actual: Negative	1,501	4,9

	Predicted as Positive	Predicted as Negativ
Actual: Positive	86.03%	13.9
Actual: Negative	23.33%	76.(

Ver 1

- Better at **correctly identifying true positives** but at the cost of a higher False Positive Rate.
- If we want to minimize missed positives (i.e., **ensure that all true positives are detected, even if it means more false alarms**), Version 1 is better.

Ver 2 - Similarity 80 - Both Scam and Duplicated		
	Predicted as Positive	Predicted as Negative
Actual: Positive	796	78(
Actual: Negative	253	6,18

	Predicted as Positive	Predicted as Negative
Actual: Positive	50.32%	49.68
Actual: Negative	3.93%	96.0

Ver 2 - Similarity 80%

- Better at **correctly identifying true negatives** but at the cost of missing more true positives.
- If we want to minimize false positives (i.e., **reduce the number of false alarms, even if it means missing some true positives**), Version 2 is better.

Conclusion

We can use this tool for 2 purposes:

1. **Internal check**: Using both 2 versions to investigate since it better covers almost duplicated/ scam cases
2. **Auto block duplicated algorithm**: If we want to block employers from posting duplicated jobs, we can use Version 2 since it has a low false alarm rate

Other than that, we can still explore more to develop version 3 since the accuracy of both versions above is still not very high.

Next step

- Rule base: f1 score => Include more features.
- Separate fraud and duplicated model.