Job Waves: Surfing the Ups and Downs of the Market

CSCI1951A Data Science Spring 2025: Christopher Chen, Eric Kim, Felix Lee, Jiwon Yoo

Introduction

The COVID-19 pandemic unleashed historic disruptions across global labor markets, exposing the fragility of employment systems and reshaping how industries adapt to shocks. We wanted to gauge how public opinion via online forums interacted with ongoing market trends to ultimately see whether there would be some form of predictability in employment across different industries.

Hypotheses

Hypothesis 1: Online discussions of employment conditions exhibit statistically significant shifts in sentiment across COVID-19 periods (pre-pandemic, pandemic, and post-pandemic), as measured by Reddit data.

Hypothesis 2: Across all industries, there is no meaningful correlation between past employment data and future employment outcomes—specifically, historical employment metrics cannot reliably predict future job openings.

Hypothesis 3: The ratio of job openings to total employees increased more in certain sectors during COVID, indicating labor supply shortages, especially in essential industries like healthcare and logistics

Data

The data was taken from two different APIs: the Bureau of Labor Statistics (BLS) and Reddit. The BLS data API provided data sets on a monthly basis across six years (2018-2023) measuring number of job openings, average earnings of all employees, and number of employees for 10 sectors: **mining and logging; construction; manufacturing; trade, transportation, and utilities; information; financial activities; professional and business services; private education and health services; leisure and hospitality; and government**.

For each sector, we found the most popular Reddit forum data (posts and comments) for each period. We used Reddit's API to scrape the most upvoted posts and related comments based on keyword searches in sector-specific subreddits, filtered afterwards by time period.

Methodology

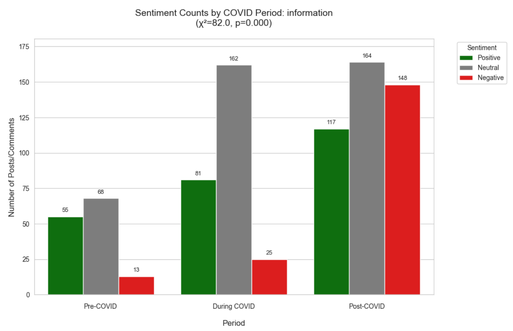
Our analysis examined 10 sectors comprising 3,679 posts/comments to identify sector-specific temporal patterns. We employed a BERT-based sentiment analyzer to classify employment-related Reddit discussions, retaining only high-relevance posts (score > 5/10). Using chi-squared tests, we compared sentiment distributions (Positive/Neutral/Negative) across three COVID-19 periods (pre/during/post; before 03/01/2020, from 03/01/2020 to 04/01/2022, and after 04/01/2022, respectively), with Cramér's V effect sizes quantifying practical significance.

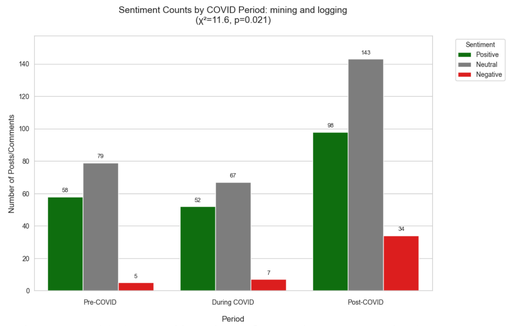
Our second analysis focused on assessing correlations between past and future data. We used XGBoost for temporal modeling across sectors, leveraging two-period lag features. Performance was assessed using normalized root mean squared error (RMSE) with 5-fold cross-validation and benchmarked against a simple lagged baseline ($X_t = X_{t-2}$).

Lastly, we calculated the job openings per employee for each sector by dividing the number of job openings by the total employees. We compared this ratio before COVID (Jan 2018–Feb 2020) and during COVID (Mar 2020–May 2022) to see if labor shortages increased. To test for significant changes, we used a two-sample t-test, which is well-suited for comparing groups with different variances.

Analysis

Claim 1





Figures 1 and 2. Grouped bar charts for sentiment counts by COVID period for the information and mining and logging sectors.

Out of ten sectors, only two (information, mining and logging) exhibited statistically significant changes in sentiment across the three COVID-19 time periods. Specifically, there is a significant increase in negative sentiment post-COVID in these two sectors.

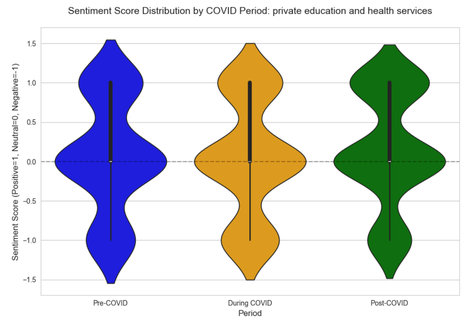


Figure 3. Violin plot indicating sentiment distribution for the private education and health services sector.

The remaining sectors did not exhibit statistically significant changes in sentiment. Most sectors had distributions similar to the one shown in the above violin plot for the private education and health services sector. Chi-squared tests confirmed the lack of significance, with all other sectors having a p-value greater than 0.05.

Claim 2

Past employment data shows meaningful predictability for future job openings, with both the machine learning model and the lag-based baseline leveraging past trends. While XGBoost achieves especially low errors in sectors like private education and professional services, gains over the baseline are modest overall, highlighting where machine learning adds sector-specific predictive value.

Sector	NRMSE (XGBoost)	Error Reduction
construction	0.026159	-0.003344
financial activities	0.027044	-0.004666
government	0.013726	-0.003149
information	0.044413	-0.007258
leisure/hospitality	0.013561	-0.00306
manufacturing	0.018172	-0.005929
mining/logging	0.118276	-0.018435
private education/health services	0.010676	-0.001887
professional/business services	0.010861	-0.00211
trade, transportation, and utilities	0.011947	-0.002901
MEAN	**0.02948**	**-0.00303**

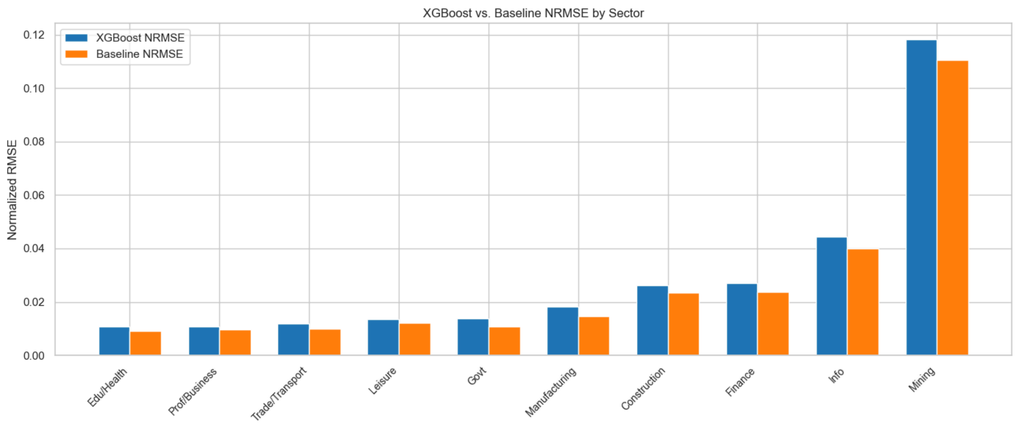


Figure 4. Grouped bar plots for each sector, comparing the XGBoost normalized RMSE with the Baseline's normalized RMSE

Claim 3

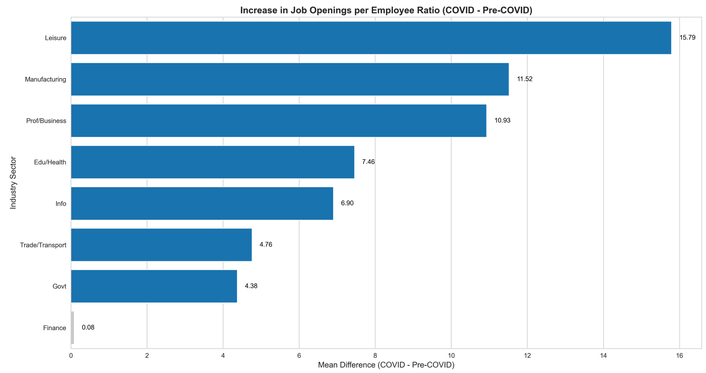


Figure 5. Bar chart showing the change in job openings per employee from Pre-COVID period to COVID period across sectors

The results show that the ratio of job openings to employees rose sharply during COVID, especially in essential sectors like leisure, healthcare, and manufacturing, signaling widespread labor shortages. In contrast, sectors like financial activities remained relatively stable, highlighting uneven pandemic impacts across industries.

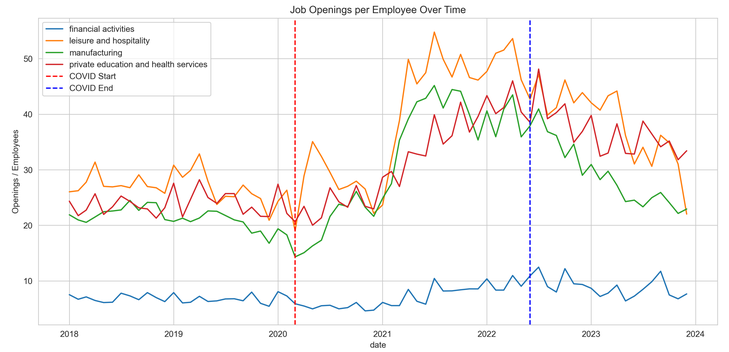


Figure 6. Line plots of how the job openings per employee ratio evolved over time in selected sectors.

Conclusion

This analysis highlights how COVID-19 reshaped labor market dynamics across industries. While online sentiment shifts were concentrated in a few key sectors, employment data revealed that historical patterns **retained predictive value**, especially when modeled with machine learning. Notably, essential and consumer-facing industries faced **sharp increases in job openings** relative to workforce size, underscoring acute labor shortages.

Challenges

- Managing data quality and consistency
  - Missing or incomplete BLS data in key sectors
  - Uneven Reddit activity across different subreddits for different industries and time periods
- Biases introduced by API limitations
  - Reddit limits the number of queries in a specific amount of time
  - Inability to filter searches to our date range, so there is no guarantee in equal amounts of data across all sectors
- Tuning sentiment analysis model to output less neutral sentiment results