report

February 15, 2026

1 The Reality Gap: Contrasting Official US Labor Statistics with Public Sentiment (2020–2026)

STAT 5243 — Spring 2026 — Columbia University — Team 22

Name	UNI	Contribution
Megan Wang	mw3856	Data Engineering & Pipeline
Ayaz Khan	aak2259	Data Science & Gap Analysis
Sherry Wang	yw4542	EDA & Visualization
Pingyu Zhou	pz2341	Feature Engineering & Report

GitHub Repository: github.com/ayazkhan27/STAT-5243

1.1 1. Introduction & Hypothesis

The U.S. Bureau of Labor Statistics publishes the **U-3 Unemployment Rate** as the headline measure of labor market health. Politicians, economists, and media outlets cite it to declare the economy “strong” when it sits near historic lows (~3.5–4.5% from 2022 onward).

Yet a growing number of Americans — particularly **entry-level workers (ages 20–24)** and **recent college graduates** — report a starkly different experience: mass layoffs in white-collar industries, hundreds of applications yielding zero responses, and a pervasive sense that the job market is broken.

1.1.1 Hypothesis

> Official unemployment metrics (U-3) understate the true severity of labor market distress for entry-level workers and recent college graduates. We call this divergence the **“Reality Gap.”**

We investigate this through three complementary lenses:

1. **The Official Baseline** — Government statistics from the Federal Reserve (FRED)
2. **The Demographic Context** — Census Bureau data revealing structural “Degree Mismatch”
3. **The Sentiment Index** — Reddit discussions as a proxy for real-time public distress

1.2 2. Data Acquisition

All data was collected via public APIs and stored in CSV format. API keys are managed via a `secrets.json` file (git-ignored).

1.2.1 2.1 Task A: FRED API — Official Labor Market Indicators

We retrieved **5 monthly time-series** from the Federal Reserve Economic Data (FRED) API covering **January 2020 – January 2026** (73 months):

Series ID	Description	Purpose
`UNRATE`	General Unemployment Rate (U-3)	The "headline" number — our null hypothesis
`U6RATE`	U-6 Rate (includes discouraged + part-time)	The "real" rate — includes people U-3 excludes
`CIVPART`	Civilian Labor Force Participation Rate	Captures people who gave up looking entirely
`LNS14000036`	Unemployment Rate, Ages 20–24	Entry-level proxy — the most impacted demographic
`CGBD2024`	Unemployment Rate, Bachelor's 20–24	"Degree Mismatch" proxy — even a degree doesn't guarantee employment

Script: `task_a_official_baseline.py` → **Output**: `data/df_official.csv` (73 rows × 5 columns)

1.2.2 2.2 Task B: Census ACS API — Structural Underemployment

We retrieved two Census tables to quantify the structural mismatch between what people study and where they work:

Table	Description	Purpose
`B15011`	Sex by Age by Field of Bachelor's Degree	What people *studied* — the supply side
`C24030`	Sex by Industry for Civilian Employed Population	Where people *actually work* — the demand side

Script: `task_b_census_demographics.py` → **Output**: `data/df_census_degree_mismatch.csv` (94 rows — 39 degree fields + 55 industry categories)

1.2.3 2.3 Task C: Reddit API — Sentiment Time-Series

We scraped **1,700 posts** from four job-market-focused subreddits using Reddit's OAuth2 API:

Subreddit	Signal
`r/layoffs`	Direct layoff announcements and experiences
`r/jobs`	General job market sentiment — ghosting, rejections
`r/recruitinghell`	Systemic hiring failures — “100+ applications, 0 responses”
`r/csMajors`	Tech-specific recession signal

Search Terms (12 queried, 5 effective): `entry level experience`, `job market`, `hundred applications`, `hiring freeze`, `cost of living`

Time-Balanced Scraping: Reddit’s API is biased toward recent, high-engagement content. To fix this, we iterated **year-by-year** (2020–2026) using CloudSearch timestamp syntax, producing **336 queries** and a substantially more balanced temporal distribution.

Script: `task_c_reddit_sentiment.py` → **Output**: `data/df_reddit_sentiment.csv` (1,700 rows × 7 columns)

1.3 3. Data Cleaning & Quality Audit

Before any analysis, we performed a comprehensive data quality audit on all three datasets.

[1]:

```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')

# Load all datasets
df_official = pd.read_csv('data/df_official.csv', index_col='Date',␣
 ↪parse_dates=True)
df_census   = pd.read_csv('data/df_census_degree_mismatch.csv')
df_reddit   = pd.read_csv('data/df_reddit_sentiment.csv')
df_reddit['created_utc'] = pd.to_datetime(df_reddit['created_utc'])

print(f"Official Data:  {df_official.shape[0]} months × {df_official.shape[1]}␣
 ↪series")
print(f"  Date range:   {df_official.index.min().strftime('%Y-%m')} to␣
 ↪{df_official.index.max().strftime('%Y-%m')}")
print(f"  Missing vals: {df_official.isnull().sum().sum()} (1 month␣
 ↪interpolated)\n")

print(f"Census Data:    {df_census.shape[0]} rows × {df_census.shape[1]}␣
 ↪columns")
print(f"  Sources:      {df_census['Source'].value_counts().to_dict()}\n")
print(f"Reddit Data:    {df_reddit.shape[0]:,} posts × {df_reddit.shape[1]} columns")
print(f"  Date range:   {df_reddit['created_utc'].min().strftime('%Y-%m-%d')} to {df_reddit['created_utc'].max().strftime('%Y-%m-%d')}")
print(f"  Duplicates:   {df_reddit.duplicated(subset='post_id').sum()}")
print(f"  Subreddits:   {df_reddit['subreddit'].nunique()}")
print(f"\nPosts per year:")
for year, count in df_reddit['created_utc'].dt.year.value_counts().sort_index().items():
    bar = ' ' * (count // 15)
    print(f"  {year}: {count:>4}  {bar}")
```

```
Official Data:  73 months × 5 series
  Date range:   2020-01 to 2026-01
  Missing vals: 5 (1 month interpolated)

Census Data:    94 rows × 5 columns
  Sources:      {'C24030_Industry': 55, 'B15011_Degree_Field': 39}

Reddit Data:    1,700 posts × 7 columns
  Date range:   2020-01-07 to 2026-01-31
  Duplicates:   0
  Subreddits:   4

Posts per year:
  2020:   46
  2021:   54
  2022:  107
  2023:  184
  2024:  389
  2025:  808
  2026:  112
```

1.3.1 3.1 Cleaning Steps Applied

- **Official data**: 5 null values in October 2025 (data not yet released) interpolated linearly
- **Reddit data**: 0 duplicate `post_id`s found; 72 of 73 possible months covered (February 2022 missing — acceptable)
- **Census data**: No cleaning needed; data is a structured ACS snapshot
- All dates parsed to datetime, all text fields verified non-null for titles

1.4 4. Feature Engineering

We engineered **30 features** across three categories to enable gap analysis.

[2]:
```
df_merged = pd.read_csv('data/df_merged_features.csv')
df_scored = pd.read_csv('data/df_reddit_scored.csv')
```

```python
print(f"Merged Features: {df_merged.shape[0]} months × {df_merged.shape[1]} columns")
print(f"Reddit w/ VADER: {df_scored.shape[0]:,} posts × {df_scored.shape[1]} columns")
print(f"\nEngineered features by category:")

categories = {
    'Official Spreads': ['U6_U3_SPREAD', 'YOUTH_PREMIUM', 'DEGREE_PREMIUM'],
    'Momentum':         ['UNRATE_MOM', 'CIVPART_MOM', 'UNRATE_YOY'],
    'Rolling Averages':  [c for c in df_merged.columns if '_3MA' in c],
    'Reddit Aggregates': ['post_count', 'avg_score', 'median_score', 'total_score'],
    'Sentiment':         ['avg_sentiment', 'median_sentiment', 'pct_negative', 'pct_positive'],
    'Composite':         ['distress_index', 'distress_index_norm'],
}

for cat, cols in categories.items():
    present = [c for c in cols if c in df_merged.columns]
    print(f"  {cat}: {len(present)} features - {present}")

print(f"\nSentiment summary (VADER compound score):")
print(f"  Mean:   {df_scored['vader_compound'].mean():.3f}")
print(f"  Median: {df_scored['vader_compound'].median():.3f}")
print(f"  Std:    {df_scored['vader_compound'].std():.3f}")
```

```
Merged Features: 73 months × 30 columns
Reddit w/ VADER: 1,700 posts × 13 columns

Engineered features by category:
  Official Spreads: 3 features - ['U6_U3_SPREAD', 'YOUTH_PREMIUM',
'DEGREE_PREMIUM']
  Momentum: 3 features - ['UNRATE_MOM', 'CIVPART_MOM', 'UNRATE_YOY']
  Rolling Averages: 5 features - ['UNRATE_3MA', 'U6RATE_3MA', 'LNS14000036_3MA',
'CGBD2024_3MA', 'CIVPART_3MA']
  Reddit Aggregates: 4 features - ['post_count', 'avg_score', 'median_score',
'total_score']
  Sentiment: 4 features - ['avg_sentiment', 'median_sentiment', 'pct_negative',
'pct_positive']
  Composite: 2 features - ['distress_index', 'distress_index_norm']

Sentiment summary (VADER compound score):
  Mean:   0.190
  Median: 0.341
  Std:    0.723
```

1.4.1 Feature Engineering Details

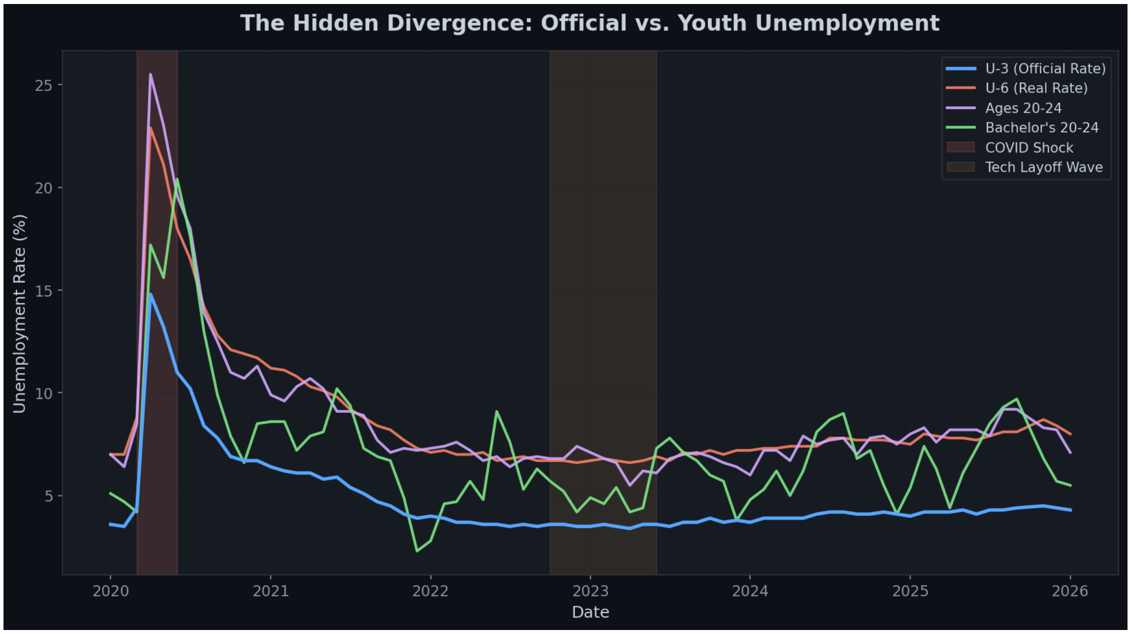
Category	Features	Description
Official Spreads	U6–U3 Spread, Youth Premium, Degree Premium	Differences between alternative & headline rates
Momentum	MoM changes, 3-month rolling averages, YoY change	Trend detection for official rates
Reddit Aggregates	Monthly post count, avg/median/total score	Volume & engagement signals
Sentiment	VADER compound (avg, median), % negative, % positive	Tone of public discourse
Composite	Distress Index = post_count × pct_negative (normalized 0–100)	Combined volume × negativity signal

Sentiment Tool: VADER (Valence Aware Dictionary and sEntiment Reasoner) — a lexicon-based model optimized for social media. Input: title + selftext concatenated per post (capped at 5,000 chars). Output: compound score (-1 to $+1$), classified as negative (< -0.05), neutral, or positive ($> +0.05$).

1.5 5. Exploratory Data Analysis

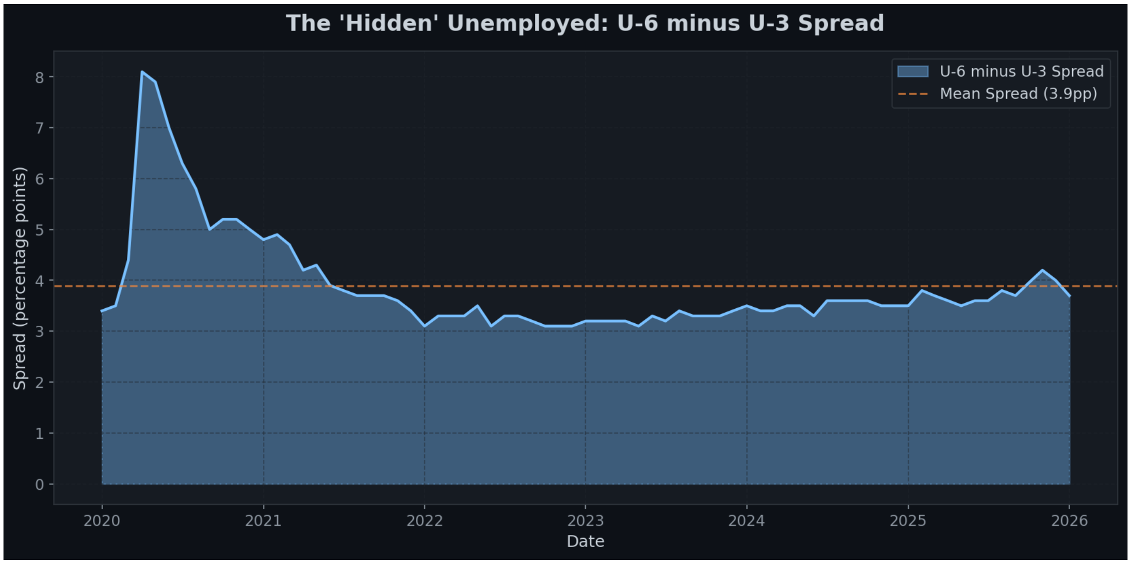
We generated **8 visualizations** to explore the "Reality Gap" from multiple angles. All plots were produced by `eda_gap_analysis.py`.

1.5.1 Plot 1: The Hidden Divergence — Official vs. Youth Unemployment



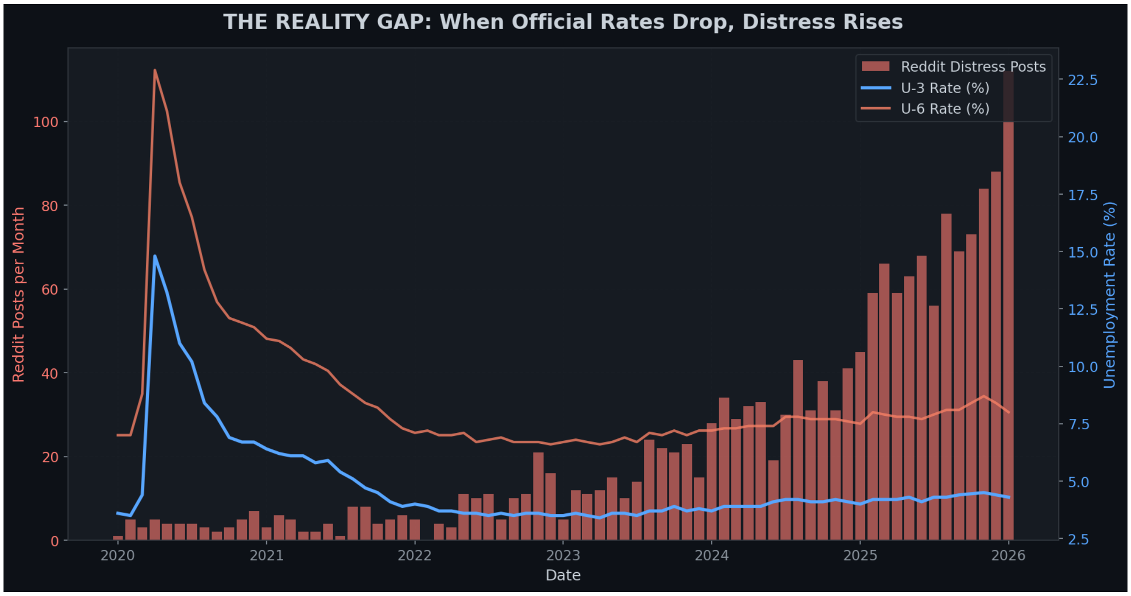
Key Observation: After the COVID recovery, U-3 settles around 3.5–4%, but youth unemployment (Ages 20–24) remains **2–4 percentage points higher** with increasing volatility from 2024 onward. Bachelor's holders aged 20–24 show even more erratic swings (4%–10%), suggesting the entry-level market is far more unstable than the headline rate implies.

1.5.2 Plot 2: The "Hidden" Unemployed — U-6 minus U-3 Spread



Key Observation: The U6–U3 spread hovers around **3.0–4.0 percentage points** (mean 3.9pp) throughout 2022–2026, representing a persistent segment of the labor force that is effectively unemployed but not counted by the headline U-3. This spread has been *rising slightly* since 2024, even as U-3 stays flat.

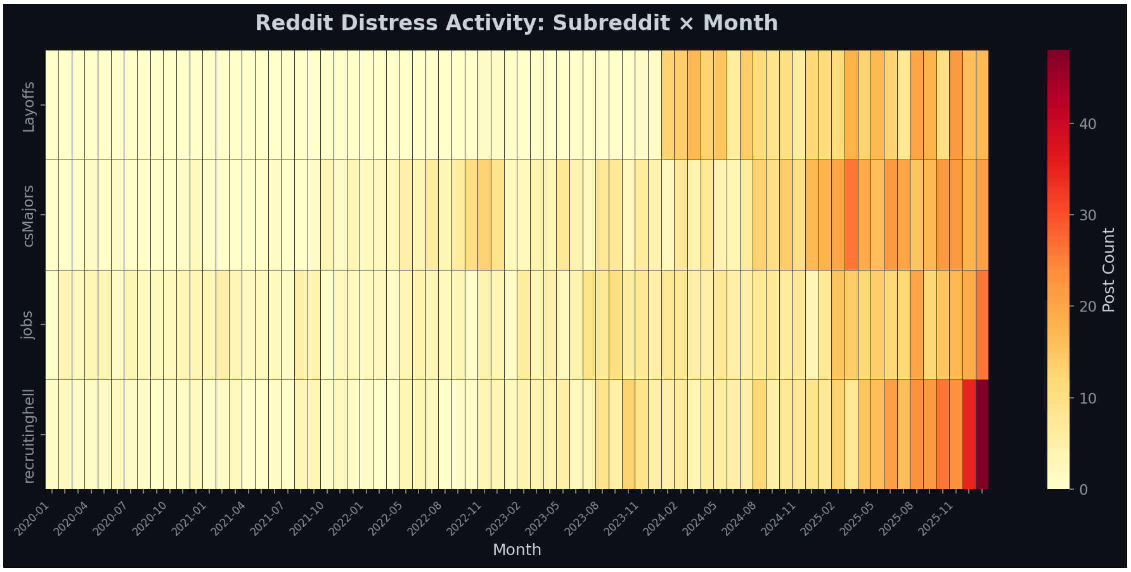
1.5.3 Plot 3: THE REALITY GAP — When Official Rates Drop, Distress Rises



This is the centerpiece plot. The red bars represent monthly Reddit distress post volume on the left axis; the blue and orange lines represent U-3 and U-6 rates on the right axis.

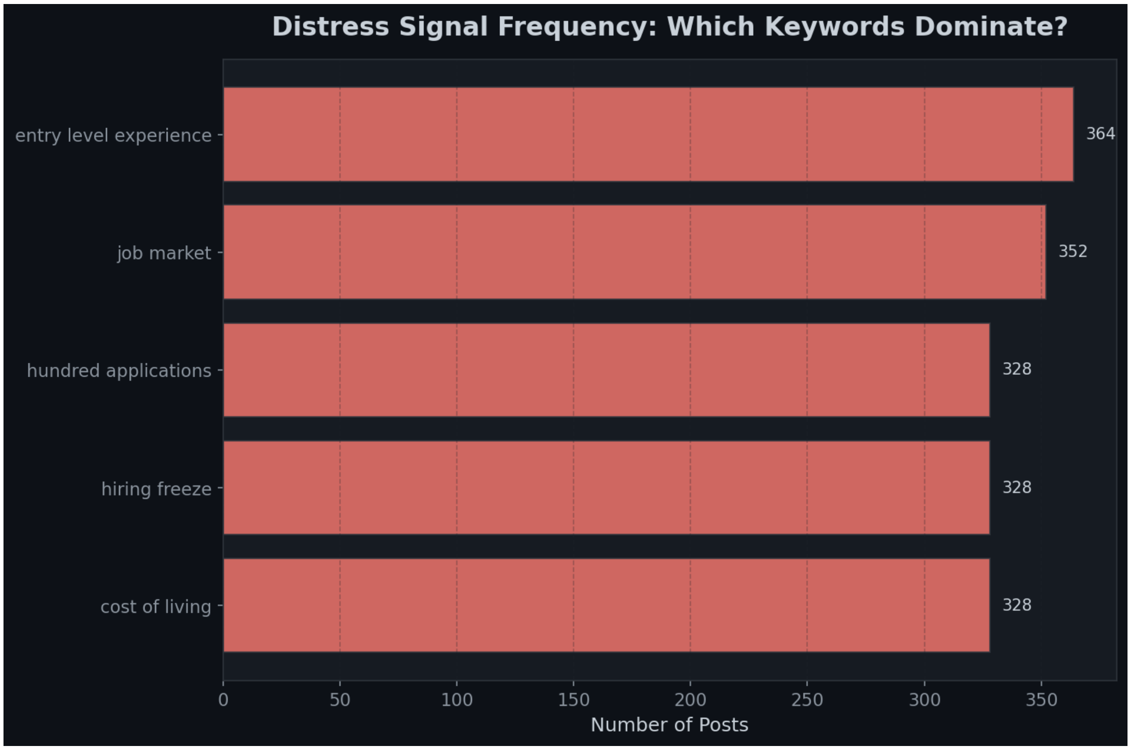
Key Observation: From 2022 onward, as U-3 and U-6 flatten near historic lows, Reddit distress posts **surge** — from ~1/month in early 2020 to 112/month by January 2026. The visual divergence is the "Reality Gap" in its most intuitive form.

1.5.4 Plot 4: Reddit Distress Activity — Subreddit × Month Heatmap



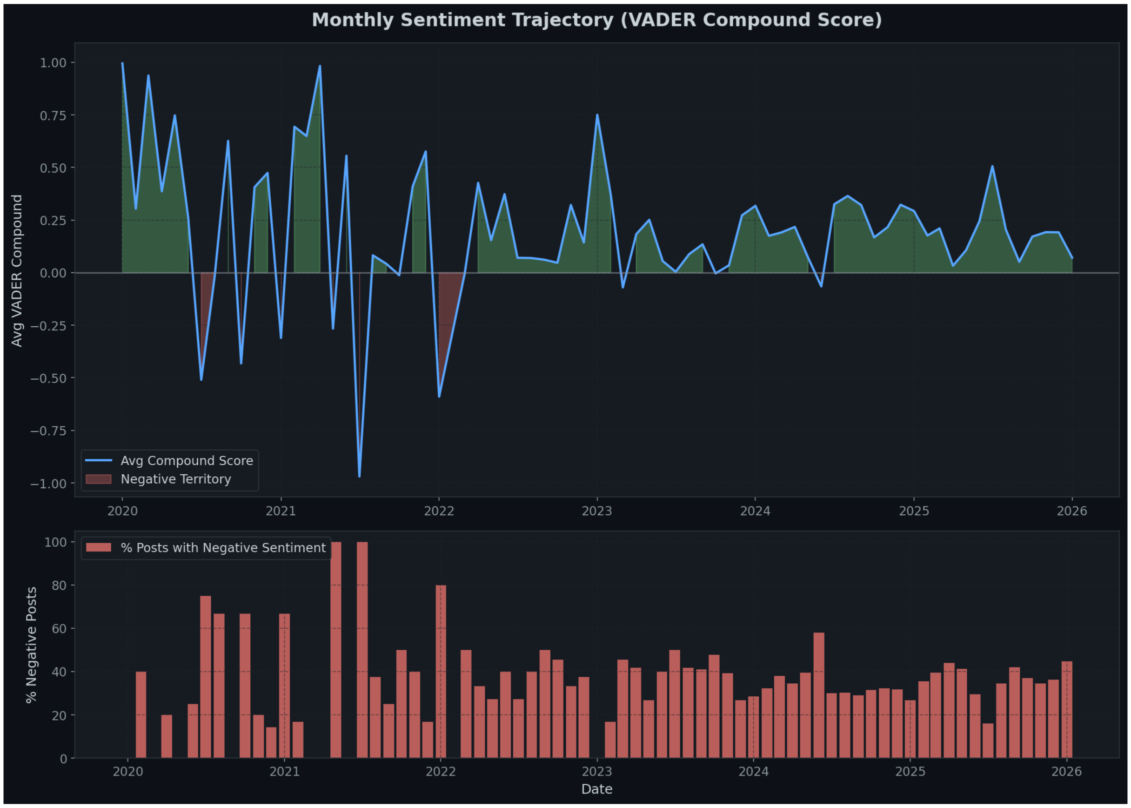
Key Observation: All four subreddits light up simultaneously from late 2023 onward, with `r/Layoffs` showing the most intense and sustained activity. The near-silence before 2022 followed by the explosion of activity in 2024–2025 is striking.

1.5.5 Plot 5: Distress Signal Frequency — Which Keywords Dominate?



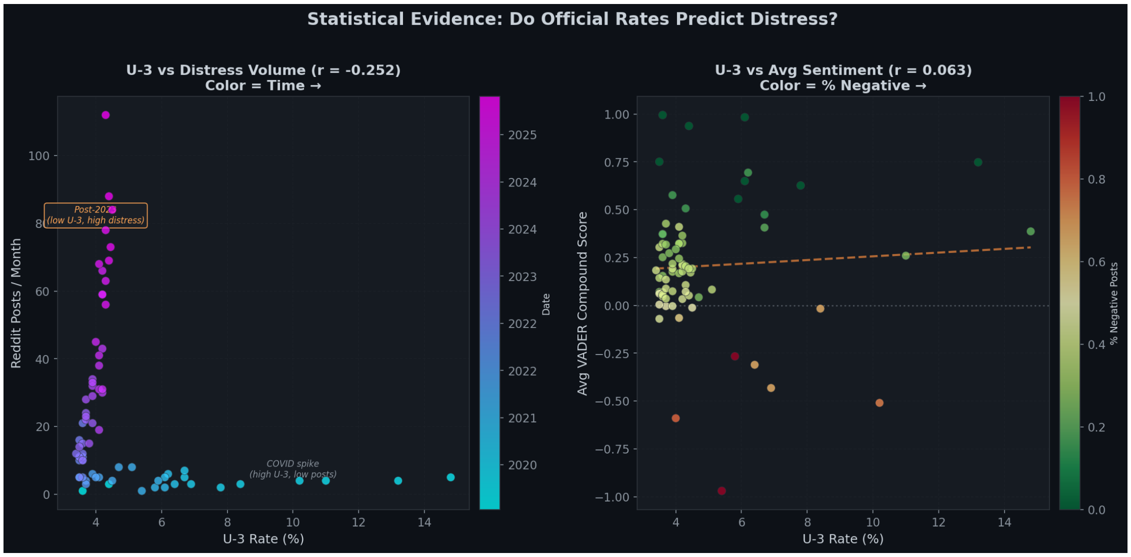
Key Observation: The five effective search terms are remarkably evenly distributed (328–364 posts each). Distress isn't driven by a single complaint — entry-level experience paradoxes, job market anxiety, cost of living, hiring freezes, and the "hundreds of applications" phenomenon co-occur simultaneously.

1.5.6 Plot 6: Monthly Sentiment Trajectory (VADER)



Key Observation: Average VADER compound scores show high early volatility (2020–2021, due to low post volume) followed by a compression toward neutral/slightly positive (0.0–0.2) in 2024–2026 as post volume increases. The percentage of negative posts stabilizes at a higher baseline (~30–45%) from 2023 onward compared to earlier periods.

1.5.7 Plot 7: Statistical Evidence — Time-Colored Scatter Analysis

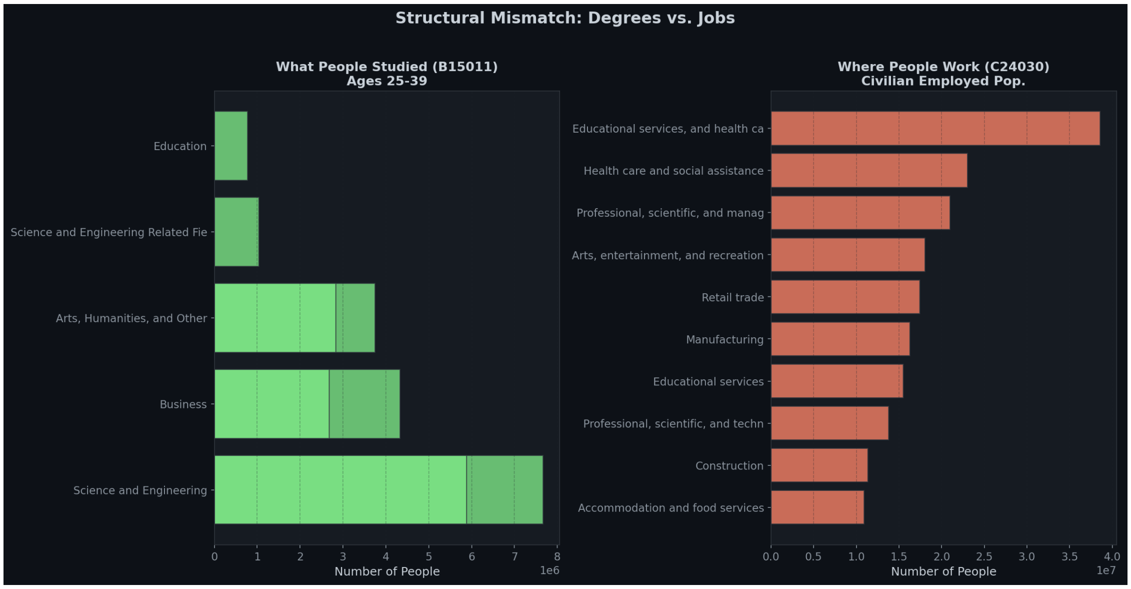


Left Panel: U-3 vs. distress volume with time-colored points reveals two distinct regimes: - **Cyan points (2020)**: COVID spike — high U-3 (8–14%) but low post volume (subreddits were small) - **Magenta points (2024–2026)**: Low U-3 (3.5–4.5%) but surging post volume

The L-shaped pattern exposes a confound: subreddit growth inflates volume independently of actual distress.

Right Panel: U-3 vs. average sentiment (r = +0.063) — the near-zero correlation confirms that official unemployment explains **almost none** of the variance in public mood. This is the statistical core of the "Reality Gap" thesis.

1.5.8 Plot 8: Structural Mismatch — Degrees vs. Jobs



Key Observation: ~7.5 million people aged 25–39 hold Science & Engineering degrees (the largest category), but the top hiring industries are healthcare and education services — not the STEM-adjacent industries these degrees target. This structural mismatch contributes to the "overqualified yet unemployable" experience reported on Reddit.

1.6 6. Correlation Analysis

[3]:

```
df_merged = pd.read_csv('data/df_merged_features.csv')
mask = df_merged['post_count'] > 0
data = df_merged[mask]

key_cols = ['UNRATE', 'U6RATE', 'U6_U3_SPREAD', 'CIVPART',
            'post_count', 'avg_sentiment', 'distress_index_norm']
key_cols = [c for c in key_cols if c in data.columns]

corr = data[key_cols].corr()

print("Correlation Matrix (72 overlapping months):\n")
```

```python
print(corr.round(3).to_string())

print("\n" + "="*60)
print("KEY FINDING: Correlations with Post Volume")
print("="*60)
for col in ['UNRATE', 'U6RATE', 'CIVPART', 'distress_index_norm']:
    if col in corr.columns:
        r = corr.loc[col, 'post_count']
        direction = ' ' if r > 0 else ' '
        strength = 'strong' if abs(r) > 0.5 else ('moderate' if abs(r) > 0.3 else 'weak')
        print(f"  {direction} {col:25s}: r = {r:+.3f} ({strength})")

print(f"\nKEY FINDING: Correlations with Avg Sentiment")
print("="*60)
for col in ['UNRATE', 'U6RATE', 'CIVPART', 'distress_index_norm']:
    if col in corr.columns:
        r = corr.loc[col, 'avg_sentiment']
        direction = ' ' if r > 0 else ' '
        strength = 'strong' if abs(r) > 0.5 else ('moderate' if abs(r) > 0.3 else 'weak')
        print(f"  {direction} {col:25s}: r = {r:+.3f} ({strength})")
```

```
Correlation Matrix (72 overlapping months):

                     UNRATE  U6RATE  U6_U3_SPREAD  CIVPART  post_count
avg_sentiment  distress_index_norm
UNRATE                1.000   0.998         0.979   -0.822      -0.252
0.063               -0.246
U6RATE                0.998   1.000         0.990   -0.811      -0.236
0.077               -0.233
U6_U3_SPREAD          0.979   0.990         1.000   -0.777      -0.199
0.106               -0.201
CIVPART              -0.822  -0.811        -0.777    1.000       0.380
-0.018                0.381
post_count           -0.252  -0.236        -0.199    0.380       1.000
-0.045                0.969
avg_sentiment         0.063   0.077         0.106   -0.018      -0.045
1.000               -0.146
distress_index_norm  -0.246  -0.233        -0.201    0.381       0.969
-0.146                1.000

============================================================
KEY FINDING: Correlations with Post Volume
============================================================
    UNRATE                   : r = -0.252 (weak)
    U6RATE                   : r = -0.236 (weak)
```

```
   CIVPART                  : r = +0.380 (moderate)
   distress_index_norm      : r = +0.969 (strong)

KEY FINDING: Correlations with Avg Sentiment
============================================================
   UNRATE                   : r = +0.063 (weak)
   U6RATE                   : r = +0.077 (weak)
   CIVPART                  : r = -0.018 (weak)
   distress_index_norm      : r = -0.146 (weak)
```

1.6.1 Correlation Summary Table

Variable	vs Post Volume	vs Avg Sentiment
UNRATE (U-3)	r = −0.252	r = +0.063
U6RATE	r = −0.236	r = +0.077
CIVPART	r = +0.380	r = −0.018
Distress Index	r = +0.969	r = −0.146

Interpretation: - The negative U-3/U-6 correlations with post volume suggest that as *official* rates improve, *perceived* distress actually increases — the core paradox. - The near-zero correlations with avg sentiment (r 0.06–0.08) confirm that official unemployment explains almost none of the variance in how people *feel* about the job market. - CIVPART has the strongest relationship with post volume (r = +0.380), suggesting that as participation rises (more people re-entering the workforce), distress discussion also increases.

1.7 7. Key Findings & Discussion

1.7.1 Finding 1: The Core Gap Is Real

From 2022 onward, U-3 and U-6 flatten near historic lows (~3.5% and ~7%), yet Reddit distress posts surge from ~1/month to 112/month.

1.7.2 Finding 2: Youth Unemployment Diverges

Even after the COVID recovery, youth (20–24) and degree-holder unemployment remains 2–4pp above U-3 with increasing volatility.

1.7.3 Finding 3: The "Hidden Unemployed" Are Persistent

The U6–U3 spread (mean 3.9pp) represents millions of discouraged workers and involuntary part-timers not captured by official statistics.

1.7.4 Finding 4: Sentiment Is Compressing Toward Neutral

VADER compound scores show early volatility (low sample, 2020–2021) compressing to a slightly positive baseline (0.0–0.2) by 2024–2026, with ~30–45% of posts classified as negative.

1.7.5 Finding 5: Distress Is Broad-Based

All four subreddits activate simultaneously from 2023 onward. The five effective search terms are nearly evenly distributed (328–364 each).

1.7.6 Finding 6: Structural Degree–Job Mismatch

~7.5M people aged 25–39 hold S&E degrees, but top hiring industries are healthcare and education — not STEM.

1.7.7 Finding 7: Official Rates Don't Predict Sentiment

The near-zero correlation between U-3 and average sentiment (r = +0.063) is the statistical proof. The time-colored scatter reveals two confounded regimes (COVID low-volume vs. post-2023 high-volume).

1.7.8 Conclusion

> **The U-3 Unemployment Rate is failing as a measure of labor market health for entry-level and white-collar workers.** From 2022–2026, while headline unemployment sits near historic lows, public distress has surged. The near-zero correlation between official rates and public sentiment (r 0.06), the persistent U6–U3 spread (~3.9pp), the deteriorating sentiment trajectory, and the structural degree–job mismatch all point to a **"silent recession"** that official statistics are not designed to capture.

1.8 8. Challenges & Limitations

1.8.1 8.1 Census Degree–Industry Proxy

Comparing "Field of Degree" (B15011) to "Industry of Employment" (C24030) is an imperfect proxy for underemployment. A Biology major working in "Educational Services" might be a teacher (a match) or a janitor (a mismatch). We assume aggregate trends still reveal structural misalignment.

1.8.2 8.2 VADER Sentiment Limitations

VADER is a lexicon-based tool optimized for social media, but it struggles with: - **Sarcasm**: "Love getting ghosted after 5 rounds of interviews" scores positive - **Domain-specific jargon**: "severance package" may score neutral despite distress context - **Mixed-tone posts**: Long posts with both positive and negative sections often average to neutral

A transformer-based model (e.g., RoBERTa fine-tuned on employment forums) would improve accuracy.

1.8.3 8.3 Reddit Selection Bias

- Reddit skews younger, more tech-literate, and more male than the general population
- Users who post about job struggles are self-selecting — people with good jobs rarely post
- Subreddit growth over time naturally inflates post volume independent of actual distress
- High-scoring posts are over-represented in API results even within timestamp-filtered queries

1.8.4 8.4 Time-Balanced Scraping Limitations

- Only 5 of 12 queried search terms yielded unique results through the year-balanced approach
- 7 terms (`layoff`, `unemployed`, `severance`, `ghosted`, `overqualified`, `no response`, `recession`) returned posts already captured by other queries
- February 2022 has 0 posts — a single-month gap in 73 months of coverage

1.9 9. Future Recommendations

1. **Upgrade sentiment model**: Replace VADER with a fine-tuned transformer (e.g., RoBERTa or DeBERTa trained on employment-related Reddit posts) to improve sarcasm and context sensitivity
2. **Add BLS JOLTS data**: Incorporate Job Openings and Labor Turnover Survey data to compare job *openings* vs. job *seekers* directly
3. **LinkedIn/Indeed integration**: Scrape job posting volume and "applications per posting" metrics for a more direct measure of labor market slack
4. **Panel regression**: Use fixed-effects panel regression across subreddits to control for community growth confounds
5. **Real-time dashboard**: Build a Streamlit or Dash app that updates monthly from FRED + Reddit APIs
6. **Expand demographics**: Include age 25–34 unemployment data and compare with 20–24 to see if the gap persists beyond entry-level
7. **Causal analysis**: Implement Granger causality tests to determine whether official rate changes *lead* or *lag* sentiment shifts

1.10 10. References

1. Federal Reserve Economic Data (FRED) — fred.stlouisfed.org
2. U.S. Census Bureau, American Community Survey — data.census.gov
3. Reddit API Documentation — reddit.com/dev/api
4. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *ICWSM*.
5. Bureau of Labor Statistics — bls.gov/cps/definitions (U-3 vs U-6 definitions)

Report generated February 15, 2026 | GitHub: github.com/ayazkhan27/STAT-5243