FLOUR: Fluctuations in Labor-market Outcomes & Unemployment Related-sentiment

Evan Chen
University of Illinois Urbana-Champaign
Champaign, IL, USA
echen48@illinois.edu

Emily Ho
University of Illinois Urbana-Champaign
Champaign, IL, USA
eeho2@illinois.edu

Rachel Tin
University of Illinois Urbana-Champaign
Champaign, IL, USA
rtin2@illinois.edu

Ethan Zhang
University of Illinois Urbana-Champaign
Champaign, IL, USA
ethanhz2@illinois.edu

Abstract

Economic shifts disproportionately impact the labor market experiences and wellbeing of young adults. This vulnerability is increasingly reflected in online public discourse, and modern social media provides unprecedented access to public sentiment and discourse on economic conditions. In this work, we investigate how sentiment patterns in sector-specific online Reddit communities (subreddits) reflect and respond to national monthly unemployment rate fluctuations. We collect a comprehensive timestamped dataset of over 400,000 Reddit posts via API. Then, using time-series analysis, Granger causality, topic modeling, and autoregressive modeling, we examine potential relationships between social media discourse and labor market trends across five sectors: Information, Education & Health, Leisure & Hospitality, Financial, and Agriculture. While significant findings were limited, likely constrained by the coarse granularity of our unemployment data, the Education & Health sector exhibited a strong inverse correlation, where rising unemployment corresponded to a decrease in negative sentiment. More importantly, we find that Education & Health subreddit discourse significantly predicts future unemployment rates. We also identified the reverse relationship in other sectors, where unemployment significantly predicts subsequent sentiment in the Leisure & Hospitality sector and money-related discourse in the Information sector. Our methodology establishes a framework for analyzing these complex, bidirectional socio-economic dynamics, highlighting online discourse as a potential leading indicator for labor market shifts.

CCS Concepts

• **Human-centered computing** → **Social media**; • **Applied computing** → **Economics**.

Keywords

Employment, Wellbeing, Labor Market Fluctuations, Sentiment

1 Introduction

The connection between economic conditions and public discourse has taken on new meanings, particularly in the wake of the digital age, where social media sites serve as real-time records of public opinion and behavioral responses to widespread economic upheavals such as the COVID-19 pandemic or trade tariffs. In contrast to traditional surveys or interviews that take time and are expensive, or are restricted by the bias of recall, digital platforms have provided access to the ways in which people understand and respond to economic shifts at speed and scale.

Economic studies have shown the psychological impact of unemployment for some time, showing that loss of jobs is associated with increased stress, anxiety, and alterations in social behavior [4, 5, 16]. These works tend to rely on surveys, clinical interviews, or longitudinal studies to provide evidence that psychological and behavioral responses can be examined, especially with regard to employment conditions in the online discourse. Social media provides a pathway for us to capture genuine and unvarnished emotional responses to employment market fluctuations and use the data as an additional source to understand how labor market conditions affect public wellbeing and whether public sentiment can predict markets.

In recent years, studies on platforms — like Twitter, Weibo, and Reddit — have documented that unemployment is associated with observable changes in language style, sentiment expressions, and behavior in an online context [10, 11, 15, 16]. This demonstrates the potential for social media to be used as an alternative method to assess labor market conditions and their social implications. Much of the previous research has either examined unemployment at the national or regional level or examined a particular group of individuals, such as university students or white collar workers. The channels through which employment conditions may influence sector-level online discussions have received little attention.

This study attempts to fill this gap by examining the relationship between unemployment rates and discussion behaviors in Reddit communities across five employment sectors. Towards this goal, we investigate the following research questions:

RQ1: Is there a significant relationship between unemployment and sector-specific Reddit community linguistic features and sentiment over time? Does it exhibit causality in either direction?

RQ2: Do significant patterns, topics, or themes in sector-specific Reddit communities emerge over time in relation to labor market fluctuations?

CS598: CWB, University of Illinois Urbana-Champaign
2025.

2 Related Work

Researchers are increasingly examining the relationships between economic situations, employment, and public wellbeing in terms of information distributed through social media. Earlier work demonstrated that changes in digital forms of behavior can tell us about existing socioeconomic inequality. For example, Llorente et al. (2015) [11] conducted research reconstructing unemployment incidence in Spain using geolocated Twitter activity, finding that temporal rhythms of activity, grammatical correctness, and diversity of movements were significantly correlated with regional unemployment. This work served as an early proof-of-concept that digital traces can serve as low-cost economic indicators that are able to outperform traditional survey-based data in terms of time and scale.

Tan et al. (2024) [16] built on this, exploring the psychological effects of unemployment on recent university graduates in China through Sina Weibo posts. They found that unemployment is closely associated with symptoms of anxiety and linguistic indicators of affect, showing how emotional and cognitive systems respond to precarious labor situations found in social media posts. Tan et al.'s work complements broad-based economic inequalities associated with unemployment by providing insights into individual-level consequences related to mental health.

Previous research has also emphasized the refinement of methodological routes in the detection and interpretation of work-related discourse. Liu et al. (2016) [10] put forward a human-in-the-loop framework to classify job-related tweets, isolating business accounts from individual users, while bringing an ethnographic and statistical angle in this framework. Liu et al. concluded that while focused job-related tweets are limited in their relationships, affective analysis yielded specific moments of data and insight on how workers reflect on their job and workplace. The authors suggest the affective analysis results provide a new dimension of identifying and recognizing labor market sentiment.

On an organizational level, An et al. (2023) [4] investigated job insecurity, stress, and anxiety within US white-collar workers. While their study was based upon a survey and not social media, they nonetheless reported that uncertainty in work conditions could serve to create distress in their sample, and suggested how digital engagement may mediate responses to stress and anxiety. The findings resonated with Liu et al. that situating online discourse within the context of employment and occupational health is important.

Ultimately, in Bak et al. (2022) [5], they examined the value of social media to engage in real-time monitoring of public mental health during periods of crisis. In particular, they analyzed Reddit posts in loneliness subreddits throughout the COVID-19 pandemic and noted significant increases related to discussions on depression and changing patterns with respect to the dominant themes in conversation. Their work indicates how online platforms still hold the potential to probe changes concerning the macro socio-economic and psychological realities of workers, especially during times of societal labor market pressures and volatility.

All prior work implies social media can be an effective way to look into psychological and economic elements of work. Most of the taxonomic studies merely address workforce unemployment signals at a population-level using a social media platform (e.g., Twitter to model unemployment), or explored the intersection in a specific demographic or organizational context (e.g., graduate workforce perspective, white-collar workers). We extend this line of inquiry as we will investigate how Reddit community-level unemployment conversations track or align with official national unemployment data. Uniquely, we establish a method to collect data containing specific Reddit community threads that align with economic sectors. We employ time-series techniques, Granger causality, topic modeling, and autoregressive modeling with the objective of developing a more nuanced understanding of the relationship between economic indicators and the lived experience discourse of persons employed within the workforce of identified sectors.

3 Data

Table 1: Subreddits by Sector and Subscriber Count

Sector	Subreddits (Subscribers)
Information	r/ITCareerQuestions (400k), r/CSCareerQuestions (1.1M), r/journalism (60k), r/publishing (25k)
Education & Health	r/Teachers (627k), r/Professors (124k), r/nursing (532k), r/medicine (461k)
Leisure & Hospitality	r/KitchenConfidential (623k), r/Serverlife (218k), r/bartenders (151k), r/talesfromthefrontdesk (937k)
Financial	r/FinancialCareers (820k), r/finance (1.9M), r/accounting (468k), r/InsuranceProfessional (27k)
Agriculture	r/farming (142k), r/agriculture (33k), r/forestry (30k), r/Truckers (219k)

Our study integrates two primary data sources: (1) monthly sector-specific unemployment rates from the U.S. Bureau of Labor Statistics (BLS), and (2) timestamped Reddit posts from occupation-related subreddits spanning August 2023 to August 2025.

3.0.1 Unemployment Data. We collected monthly, industry-specific unemployment rates from the Federal Reserve Economic Data (FRED) database maintained by the Federal Reserve Bank of St. Louis [1]. The BLS data portal could not provide the month-by-month granularity required for our analysis, requiring manual data collection from FRED for each of the 25 months in our study period. The resulting data includes unemployment rates for five sectors: Information, Financial Activities, Education & Health, Leisure & Hospitality, and Agriculture.

3.0.2 Reddit Data Collection. For each sector, we identified representative subreddits based on two criteria: (1) alignment with prominent occupations within the sector according to BLS North American Industry Classification System (NAICS) categories [2], and (2) minimum membership of 20,000 subscribers to ensure sufficient post volume for analysis. Membership counts were obtained from Subreddit Stats [3]. Table 1 details the selected subreddits and their respective subscriber counts. We utilized the Arctic Shift API

Table 2: LIWC Category Descriptive Statistics by Sector (Mean ± σ)

Sector	Positive Emotion	Negative Emotion	Achievement	Work	Money
Information	5.520 ± 0.066	0.723 ± 0.033	1.854 ± 0.033	4.755 ± 0.095	0.757 ± 0.024
Education & Health	5.536 ± 0.045	1.159 ± 0.072	1.515 ± 0.043	4.079 ± 0.068	0.501 ± 0.031
Leisure & Hospitality	5.280 ± 0.101	1.332 ± 0.064	1.490 ± 0.049	2.193 ± 0.083	0.818 ± 0.078
Financial	5.252 ± 0.070	0.807 ± 0.027	1.616 ± 0.033	4.146 ± 0.082	2.205 ± 0.064
Agriculture	5.203 ± 0.148	1.146 ± 0.116	1.269 ± 0.068	2.249 ± 0.123	0.848 ± 0.075

[9], an open-source Reddit data aggregator, to overcome Reddit's native API rate limits. Our collection script (Appendix A) retrieved all posts recursively in batches from August 2023 through August 2025. Posts were initially stored in separate CSV files for each unique subreddit-year-month combination.

3.0.3 Data Preprocessing. We performed light text preprocessing including: (1) lowercasing all text, (2) removing posts marked as deleted, removed, or empty, and (3) aggregating posts by sector and month to align with unemployment data granularity. Stopword removal was selectively applied—retained for LIWC and sentiment analysis to preserve linguistic patterns, but removed for BERTopic modeling to enhance topic coherence.

After preprocessing, our dataset comprises 411,396 posts distributed across five sectors: Education & Health (131,979 posts), Financial (117,056), Information (70,385), Leisure & Hospitality (56,301), and Agriculture (35,675). The cleaned data is publicly available in our GitHub repository in `/data/cleaned`.

3.1 Descriptive Statistics

Table 3: VADER Sentiment Descriptive Statistics by Sector

Sector	Negative		Positive		Compound	
	M	σ	M	σ	M	σ
Information	0.066	0.002	0.188	0.003	0.539	0.013
Education & Health	0.099	0.005	0.174	0.002	0.298	0.031
Leisure & Hospitality	0.102	0.003	0.172	0.003	0.263	0.016
Financial	0.082	0.002	0.165	0.002	0.315	0.011
Agriculture	0.088	0.005	0.157	0.004	0.229	0.027

Table 3 presents baseline sentiment characteristics across sectors using VADER (Valence Aware Dictionary and Sentiment Reasoner) scores. Notably, the Information sector stands apart with the most positive sentiment profile, evidenced by the highest mean compound (M = 0.539) and positive (M = 0.188) scores, as well as the lowest mean negative (M = 0.066) score. In contrast, the Leisure & Hospitality (M = 0.102) and Education & Health (M = 0.099) sectors exhibit the highest mean negative sentiment. It is also worth noting that the standard deviations are consistently low across all sectors ($\sigma \leq 0.005$), suggesting a high degree of sentiment homogeneity within the posts categorized for each sector.

Table 2 shows LIWC (Linguistic Inquiry and Word Count) category prevalence by sector across the entire dataset. Notably, work-related terms are most frequent in the Information (M = 4.755), Financial (M = 4.146), and Education & Health (M = 4.079) sectors, suggesting a strong thematic focus on employment in those fields. Furthermore, achievement language is highest in the Information sector (M = 1.854), potentially reflecting discourse around career advancement and technical accomplishments.

4 Methods

Our analytical approach employs four complementary methods to address our research questions: VADER sentiment analysis, LIWC psycholinguistic feature extraction, BERTopic modeling, and autoregressive modeling. For VADER and LIWC specifically, we conduct time-series statistical tests (Pearson correlation and Granger causality) to evaluate significance of results.

4.1 RQ1

To examine temporal relationships between unemployment and linguistic patterns (RQ1), we ran VADER sentiment analysis, LIWC feature extraction, and autoregressive modeling.

4.1.1 VADER Sentiment. We applied the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analyzer to all posts after removing stopwords. VADER computes four metrics per text: negative, neutral, positive, and compound scores. The compound score represents a normalized, weighted composite ranging from -1 (extremely negative) to +1 (extremely positive), making it particularly suitable for capturing overall sentiment. For each sector-month combination, we calculated mean VADER scores across all posts, producing monthly time series of sentiment indicators. This aggregation reduces noise from individual posts while preserving temporal patterns.

4.1.2 LIWC. We used Linguistic Inquiry and Word Count (LIWC) [6], a psycholinguistic lexicon that categorizes words into psychological and topical dimensions. We focused on five categories relevant to employment discourse:

- **posemo**: Positive emotion words (e.g., "happy")
- **negemo**: Negative emotion words (e.g., "anxious")
- **achiev**: Achievement-related terms (e.g., "earn," "succeed")
- **work**: Work-domain vocabulary (e.g., "job," "career")
- **money**: Financial terms (e.g., "salary," "compensation")

To improve computational efficiency, we constructed prefix-match and exact-match lookup tables from the LIWC dictionary. For each post, we tokenize the text, counted category matches, and normalized by total word count to produce percentages. Monthly averages were computed per sector. Unlike VADER, LIWC analysis retained stopwords to preserve authentic linguistic patterns.

4.1.3 Statistical Tests. We assessed two types of relationships between unemployment rates and linguistic features. For both tests, following convention, we set $\alpha = 0.05$ for significance.

- **Pearson Correlation:** We computed correlation coefficients (r) and p-values between monthly unemployment rates and each linguistic feature within sectors.
- **Granger Causality:** To test for bidirectional predictive power, we applied Granger causality tests [7] with a 1-month lag. This test was run in two directions: (1) whether past unemployment rates *predict* current linguistic patterns, and (2) whether past linguistic patterns *predict* current unemployment rates. We used the F-test variant implemented in Python's `statsmodels` library.

Using the `causal` library [14], we also attempted to measure the impact of unemployment rates on sentiment and vise versa with the event set to April 2025 (80% of the data). We choose to omit these results, as they did not yield significant causation in either direction between the unemployment rates and online sentiment.

4.1.4 Autoregression. We applied autoregressive (AR) and autoregressive with exogenous variables (ARX) models to predict monthly unemployment rates across five economic sectors using historical data from the Federal Reserve Economic Data (FRED) database. Monthly unemployment rates were merged with LIWC features to examine whether linguistic-emotional indicators could enhance predictive accuracy. Each dataset included the year, month, and a computed date variable to support time series analysis. Two models were trained for each sector. The baseline AR(1) model predicted the unemployment rate at time t based solely on its lagged value, while the ARX model extended this framework by incorporating contemporaneous LIWC features as exogenous regressors. Each model was trained on the first 80 percent of the time series data and evaluated on the remaining 20 percent to assess out-of-sample predictive performance. Model accuracy was assessed through the use of the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2), both of which access prediction error and variance explained, respectively. To assess if the use of linguistic-emotional features improved predictive accuracy, the relative change in RMSE, and increase in R^2 was assessed as a function of the difference between models AR and model ARX. Symmetric means absolute percentage error (SMAPE) was also used, but results were inconclusive. All analysis was conducted in Python using the `statsmodels` and `scikit-image` libraries.

4.2 RQ2

To identify potential themes in employment discourse, we used BERTopic [8], a neural topic modeling approach combining transformer embeddings, dimensionality reduction, and clustering.

4.2.1 BERTopic. Our pipeline consisted of four stages:

(1) **Embeddings:** We encoded post text using the Sentence-Transformer model `all-MiniLM-L6-v2`, a lightweight BERT variant. Stopwords were removed prior to encoding to improve topic coherence. Batch size was set to 64 to balance memory and speed, computation utilized an Apple M3 GPU.
(2) **Dimensionality Reduction:** UMAP (Uniform Manifold Approximation and Projection) [13] reduced embeddings to 5 dimensions using cosine distance, 15 neighbors, and minimum distance of 0. We set the random seed to 598.
(3) **Clustering:** HDBSCAN [12] clustered reduced embeddings with minimum cluster size of 50
(4) **Topic Representation:** BERTopic extracted the top 20 most characteristic words per topic using class-based TF-IDF. We specified `nr_topics=10` to limit topics per sector.

Models were trained separately per sector (5 models total) and saved using PyTorch serialization to avoid redundant computation.

4.2.2 Employment Topic Selection. Following model training, we manually reviewed the topics of each sector to identify employment-relevant topics. We examined keywords and sample posts, selecting topics with multiple keywords related to jobs, hiring, and compensation. Selected employment topic IDs (relevant keywords) were:

- **Information:** Topic 0 (job, work, experience), Topic 4 (resume, feedback), Topic 8 (salary, compensation, glassdoor)
- **Education & Health:** Topic 1 (nursing, work, job), Topic 6 (union, bonus, pay, contract)
- **Leisure & Hospitality:** No employment topics identified
- **Financial:** Topic 0 (accounting, job, work, experience)
- **Agriculture:** Topic 1 (company, work, jobs)

4.2.3 Topic Prevalence Analysis. For each selected topic, we computed monthly prevalence as the proportion of posts assigned to that topic within each sector-month. We then tested correlations and Granger causality between unemployment rates and topic prevalence using the same methods as RQ1.

4.3 Reproducibility and Code

All preprocessing scripts and method implementations are available as notebooks in our GitHub repository. Random seeds were fixed where applicable (UMAP for BERTopic). VADER and LIWC dictionaries are publicly available. The complete analytical pipeline can be replicated given our open data and documented codebase.

5 Results

We organize our results by research question, reporting statistical findings and interpretive insights when applicable. Results for all methods involving statistical tests can be viewed in Table 4.

5.1 RQ1

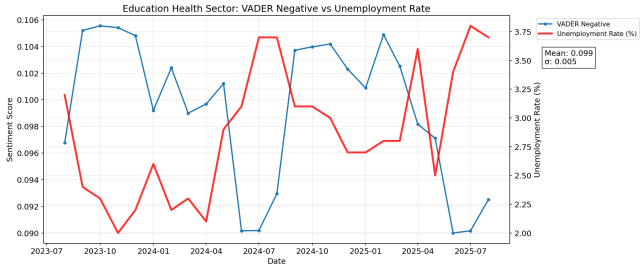

Figure 1: Education and Health - VADER Negative

5.1.1 VADER Sentiment. Conducting Pearson's correlation tests, only the Education and Health sector demonstrated significant correlations. The VADER Compound score, representing overall

Table 4: Statistical Tests of VADER, LIWC, and BERTopic Results

Method	Sector	Feature	Corr (r)	Corr p-value	Granger F†	Granger p†	Granger F‡	Granger p‡
VADER	Agriculture	Positive	0.184	0.3773	0.786	0.3852	3.003	0.0978
		Negative	-0.138	0.5102	0.084	0.7747	0.179	0.6769
		Compound	0.313	0.1280	0.152	0.7002	1.080	0.3104
	Education & Health	Positive	0.280	0.1748	0.366	0.5518	3.972	0.0594
		Negative	**-0.703**	**0.0001**	0.680	0.4190	**4.399**	**0.0482**
		Compound	**0.676**	**0.0002**	2.268	0.1470	**5.861**	**0.0246**
	Financial	Positive	-0.051	0.8093	0.075	0.7875	0.692	0.4147
		Negative	0.050	0.8133	3.026	0.0966	0.907	0.3517
		Compound	-0.105	0.6158	0.690	0.4157	2.257	0.1479
	Information	Positive	-0.193	0.3559	2.310	0.1434	0.132	0.7198
		Negative	-0.160	0.4438	3.221	0.0871	0.006	0.9382
		Compound	0.142	0.4975	0.920	0.3484	0.015	0.9028
	Leisure & Hospitality	Positive	-0.262	0.2055	0.357	0.5568	2.419	0.1348
		Negative	-0.326	0.1121	2.044	0.1675	2.475	0.1306
		Compound	0.130	0.5369	**4.937**	**0.0374**	3.466	0.0767
LIWC	Agriculture	Positive Emotion	0.238	0.2518	0.058	0.8115	0.046	0.8326
		Negative Emotion	-0.175	0.4040	0.077	0.7845	0.524	0.4773
		Achievement	-0.022	0.9167	2.022	0.1697	0.226	0.6391
		Work	0.045	0.8322	0.255	0.6189	0.308	0.5848
		Money	0.094	0.6562	1.383	0.2527	1.138	0.2982
	Education & Health	Positive Emotion	0.081	0.6994	0.393	0.5375	0.454	0.5078
		Negative Emotion	**-0.739**	**0.0000**	0.066	0.8003	2.841	0.1067
		Achievement	**0.622**	**0.0009**	4.166	0.0540	0.120	0.7326
		Work	0.303	0.1403	0.458	0.5060	0.101	0.7542
		Money	**0.589**	**0.0019**	0.002	0.9695	**4.566**	**0.0445**
	Financial	Positive Emotion	-0.148	0.4797	0.723	0.4049	0.043	0.8369
		Negative Emotion	-0.083	0.6941	1.516	0.2318	0.378	0.5454
		Achievement	-0.314	0.1263	0.108	0.7459	0.135	0.7174
		Work	0.099	0.6362	0.483	0.4947	**4.357**	**0.0492**
		Money	0.291	0.1585	0.998	0.3292	1.829	0.1906
	Information	**Positive Emotion**	-0.347	0.0896	**5.626**	**0.0273**	0.260	0.6158
		Negative Emotion	-0.165	0.4300	2.121	0.1601	0.730	0.4027
		Achievement	-0.007	0.9735	0.079	0.7812	3.748	0.0664
		Work	-0.093	0.6584	0.164	0.6900	0.265	0.6118
		Money	-0.055	0.7950	**5.113**	**0.0345**	0.420	0.5240
	Leisure & Hospitality	Positive Emotion	-0.173	0.4070	0.027	0.8715	1.696	0.2070
		Negative Emotion	0.262	0.2058	0.829	0.3728	0.109	0.7440
		Achievement	-0.320	0.1189	0.534	0.4728	0.774	0.3888
		Work	0.005	0.9799	0.660	0.4256	1.657	0.2120
		Money	-0.360	0.0774	0.029	0.8663	0.677	0.4200
BERTopic	Agriculture	**Topic 1**	-0.387	0.0560	**8.741**	**0.0075**	0.195	0.6630
	Education & Health	Topic 1	0.383	0.0587	0.645	0.4308	3.622	0.0708
		Topic 6	-0.207	0.3197	2.729	0.1134	0.047	0.8312
	Financial	Topic 0	-0.097	0.6444	0.361	0.5543	0.101	0.7539
	Information	Topic 0	-0.219	0.2939	0.478	0.4971	2.539	0.1260
		Topic 4	0.112	0.5947	0.005	0.9442	1.249	0.2763
		Topic 8	-0.389	0.0544	2.404	0.1359	1.049	0.3174

† *Unemployment rate Granger-causes Feature;* ‡ *Feature Granger-causes unemployment rate*
Bolded values indicate statistical significance at $p < 0.05$

sentiment, was found to be strongly positively correlated with the unemployment rate (r = 0.676, $p = 0.0002$). In tandem, the VADER Negative score, representing the proportion of negative sentiment, exhibited a strong negative correlation with the unemployment rate (r = −0.703, $p = 0.0001$). As seen in Figure 1, there is a clear inverse relationship between negative sentiment and the unemployment rate. This suggests that the Education and Health subreddit communities exhibit a unique phenomena: as unemployment rates increase, there is less negative discourse among community members. This observation is counterintuitive to standard economic expectations, where a rise in unemployment often correlates with negative sentiment. A possible interpretation is that the communities might shift their focus away from grievances related to employment security

and toward supportive, informational, or policy-focused discussions when the employment outlook is poor. However, this is merely an observation of correlation, not causation.

Moving beyond correlation, we conducted bidirectional causality tests. First, testing if unemployment predicts sentiment, we found results consistent with our initial analysis: only the Leisure & Hospitality sector's Compound VADER score was significantly Granger-caused by the unemployment rate ($F = 4.937$, $p = 0.0374$). The F-value of 4.937 indicates a modest but statistically significant predictive power of the unemployment rate over the sentiment expressed in Leisure & Hospitality subreddits. This suggests that changes in the unemployment rate in one month are likely to precede a change in the overall sentiment of the sector's online discourse in the following month. The stronger causal link in Leisure & Hospitality could reflect a unique nature of this sector, where economic shifts quickly translate into job security concerns or news that rapidly influences the mood of the workforce.

More notably, testing the reverse direction, we found that sentiment in the Education and Health sector significantly predicts future unemployment rates. Both the VADER Negative score ($F = 4.399$, $p = 0.0482$) and the VADER Compound score ($F = 5.861$, $p = 0.0246$) were found to Granger-cause unemployment with a one-month lag. This finding supports the hypothesis that online discourse can serve as a leading indicator for labor market shifts, particularly in this sector. The strong correlations, combined with this causality, suggest that the observed decrease in negative sentiment could be a predictive signal. Other sectors did not exhibit significant causality, as detailed in Table 4. A key limitation that may obscure causal relationships in other sectors is the granularity of the temporal data, which is only monthly. More frequent sampling (e.g., weekly or daily) could capture shorter-term causal dynamics that are averaged out in the current monthly time series.

Additional figures displaying significant results from VADER sentiment analysis are available in Appendix B.1.

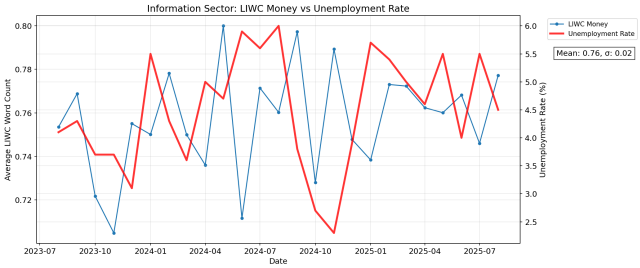

Figure 2: Information - LIWC Money

5.1.2 LIWC. The Pearson correlation analysis after LIWC revealed significant linear relationships only within the Education and Health sector. Specifically, three LIWC categories demonstrated statistically robust correlations:

- LIWC Achievement (r = 0.622, $p = 0.0009$)
- LIWC Money (r = 0.589, $p = 0.0019$)
- LIWC Negative Emotion (r = −0.739, $p < 0.0001$)

The results indicate that higher unemployment in the Education and Health sector correlates with an increase in discourse related to achievement and money, alongside a simultaneous decrease in the frequency of negative emotion words. Similar to the VADER findings, this suggests that the online communities in this sector may adopt a more proactive, goal-oriented, or even supportive tone (focusing on career/achievement and financial stability) rather than expressing increased frustration or sadness when the employment climate worsens. These observed relationships signify that the linguistic sentiments and economic trends covary: causality tests confirm this predictive relationship for some features.

Our bidirectional causality tests yields significant results in both directions, as detailed in Table 4. First, looking at unemployment predicting linguistics, the Information Sector showed significance. LIWC Positive Emotion ($F = 5.626$, $p = 0.0273$) and LIWC Money ($F = 5.113$, $p = 0.0345$) were Granger-caused by unemployment rates. These results were expected, given rising unemployment would logically precede an increase in both money-related talk and positive emotion in the subsequent month's discourse within the Information Sector's online communities. As illustrated by Figure 2, the LIWC Money word count clearly follows the unemployment rate. The predictive link between unemployment and LIWC Money makes sense, a change in the unemployment rate reasonably triggers online discussion about compensation and job market value. Higher unemployment rates Granger-causing lower LIWC Positive Emotion (and vice-versa) makes sense as communities struggle in response to perceived instability.

Second, testing whether linguistics predict unemployment, we found significant results in the Education & Health and Financial sectors. In Education & Health, the LIWC Money feature significantly Granger-caused unemployment ($F = 4.566$, $p = 0.0445$). This reinforces the VADER results, suggesting that discourse around finances in this sector acts as a leading indicator for employment shifts. Additionally, in the Financial sector, discussion of Work was found to Granger-cause unemployment ($F = 4.357$, $p = 0.0492$). Both have potential as predictors for unemployment.

Overall, the bidirectional findings suggest a complex relationship: in some sectors (Information), discourse reacts to unemployment, while in others (Education & Health, Financial), discourse appears to predict unemployment. The lack of broader causality suggests that for most sectors, either the unemployment changes do not significantly drive online discourse, or the relationship occurs at a finer temporal resolution than our monthly tests can capture.

Additional figures displaying significant results from LIWC analysis are available in Appendix B.2.

5.1.3 Autoregression. We conducted autoregression across the each sector to further investigate the relationship between unemployment rates and linguistics. We use both AR, the autoregression value that only uses past values to make prediction, and ARX, the autoregression value that uses exogenously variables to make predictions. The comparison of baseline autoregressive models to those with added linguistic/emotion features yielded mixed effects on predictive performance across sectors. For most sectors, adding LIWC features did not improve accuracy, and in many cases, resulted in a poorer fit. Specifically, the Information sector and Financial Sector model both resulted in higher RMSE and lower R^2 after adding LIWC variables, suggesting that the variables introduced noise or

overfitting rather than helping predictive performance. The Agriculture sector also exhibited a higher prediction error under the ARX model with added LIWC features, indicating that the linguistic variables were not informative to capture unemployment trends in this sector.

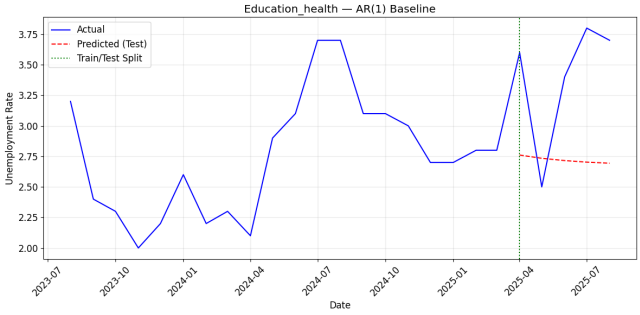

Figure 3: Education Health AR - FRED Only

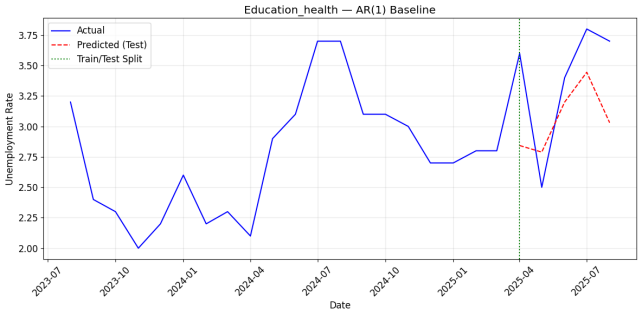

Figure 4: Education Health AR - FRED + LIWC

However, in certain sectors, particularly Education and Health, the inclusion of LIWC features substantially improved predictive performance as seen in Figure 3 and Figure 4. The ARX model reduced RMSE by approximately 39 percent and increased R^2 by nearly two points relative to the baseline AR model. This suggests that linguistic expressions of emotion, achievement, and work-related discourse may reflect underlying social or economic sentiments that correspond to fluctuations in employment within this sector. The Leisure & Hospitality sector also showed slight improvement in fit, though the magnitude of change was minimal.

Additional figures displaying results from autoregression analysis in all sectors are available in Appendix B.3.

5.2 RQ2

5.2.1 BERTopic. Our topic model was able to produce coherent topics for each sector for us to manually review. Notably, Leisure & Hospitality produced no identifiable employment topics despite prior significant observations in other methods. Looking into the topic counts, this was a direct result of BERTopic's difficulty clustering highly heterogeneous content, given that 55,983 of the 56,301 topics (99.4%) were all placed into a single topic. Following the pipeline detailed in subsubsection 4.2.1, we used monthly topic prevalence to run both correlation and Granger causality against sector-specific unemployment rates.

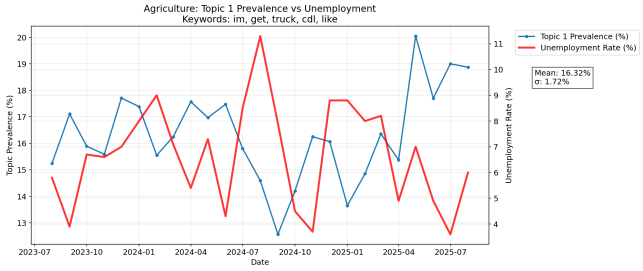

Figure 5: Agriculture Topic 1 Prevalence

The Pearson correlation analysis did not yield any correlations that met the $p < 0.05$ threshold for statistical significance. However, it is notable that several key employment topics registered p-values very close to this cutoff, suggesting potential weak relationships that may be obscured by data granularity. Specifically, Information sector's Topic 8 (keywords: salary, compensation, glassdoor) showed a moderate negative correlation approaching significance ($r = -0.389$, $p = 0.0544$). This trend, depicted in Figure 6 , suggests that as unemployment rates rise, discussion prevalence for this topic tended to decrease. This is a somewhat counterintuitive finding, indicating that online communities in this sector may discuss compensation less frequently during periods of higher unemployment.

Similarly, Agriculture Topic 1 (keywords: company, work, jobs), as shown in Figure 5 , showed a negative trend ($r = -0.387$, $p = 0.0560$). Conversely, Education and Health Topic 1 (keywords: nursing, work, job) exhibited a positive trend ($r = 0.383$, $p = 0.0587$), aligning with the LIWC findings that discourse in this sector increases with unemployment.

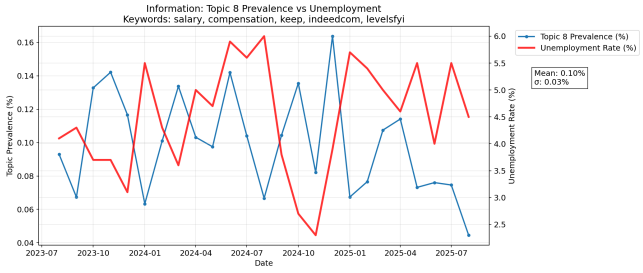

Figure 6: Information Topic 8 Prevalence

Bidirectional causality analysis identified one statistically significant predictive relationship. The prevalence of Agriculture Topic 1 was found to be significantly Granger-caused by the unemployment rate with a one-month lag ($F = 8.741$, $p = 0.0075$). This finding suggests that changes in the unemployment rate are a modest but significant predictor of the volume of discussion around jobs in the subsequent month. This makes sense, as unemployment should directly influence the amount of discussion related to job seeking and qualifications in that field. No other topics across any sector

demonstrated significant causality, as detailed in Table 4. Again, the lack of strong statistical results may be attributable to the monthly granularity of the temporal data. More immediate relationships between economic events and online discourse topics could exist.

Additional figures displaying results from BERTopic prevalence for relevant topics in all sectors are available in Appendix B.4.

6 Discussion

Our study establishes a connection between macro-level labor market indicators (monthly unemployment rates) and micro-level online discourse (sentiment, linguistic features, and topic prevalence) across five U.S. economic sectors. The frequent lack of statistical significance underscores the methodological challenge of aligning monthly labor statistics with real-time social media data. Nevertheless, the patterns identified provide valuable evidence that online sentiment and discourse can act as a proxy for socioeconomic perception that is not captured by traditional labor data.

6.1 Implications

6.1.1 Theoretical Implications. The results underscore the complexity of linking standard, delayed economic data with high-frequency online expressions in social media. The significant findings demonstrate that this relationship is highly sector-dependent and nonlinear. Correlations observed in the Education and Health sector ($r = -0.703$ for Negative VADER; $r = -0.739$ for LIWC Negative Emotion) are notable. As unemployment rises, the online discourse in these communities becomes less negative and focuses more on achievement ($r = 0.622$) and money ($r = 0.589$). This observation is counterintuitive and suggests a nuanced response, where distress may trigger proactive, positive discussion instead of frustration. Our bidirectional causality tests strongly support this. We found that sentiment and linguistic features in the Education & Health sector significantly Granger-cause future unemployment, specifically for VADER Negative ($p = 0.0482$), VADER Compound ($p = 0.0246$), and LIWC Money ($p = 0.0445$). This finding is further bolstered by the successful application of the ARX model in the Education and Health sector, which reduced the RMSE by 39% when incorporating LIWC features, supporting the view that these linguistic expressions are capturing a true underlying signal. This demonstrates relationship predicting unemployment from linguistic features. We also confirmed the expected inverse relationship in other sectors. For instance, unemployment was found to predict overall sentiment in the Leisure & Hospitality sector ($p = 0.0374$), and the positive emotion ($p = .0273$) or discussion of money ($p = 0.0345$) in the Information sector. Similarly, the prevalence of an employment-related topic in Agriculture was predicted by the unemployment rate ($p = 0.0075$). The findings that discourse *reacts* to unemployment in some sectors while *predicting* it in others supports the growing view that social media platforms can serve as valuable, complex proxies in socioeconomic sentiment, offering insight into how people collectively perceive and react to shifts in employment and economic stability.

6.1.2 Practical Implications. Given the identified predictive and correlative links between linguistic features and labor market data, analysts can potentially design models that utilize online sentiment as a supplementary, high-frequency indicator. Monitoring sector-specific linguistic features (like those found to predict unemployment from the Education and Health or Financial sector) could provide an earlier signal of workforce sentiment than slow monthly official reports. This enhanced insight could enable more meaningful and timely understanding of how labor market shifts manifest in online public conversations.

6.1.3 Ethical Implications. We acknowledge the ethical responsibilities that come with using social media data, particularly regarding privacy and representation. Future studies should carefully consider demographic and platform biases, as certain populations may be over or underrepresented online, potentially skewing analyses. Furthermore, transparency in data collection methods and algorithmic interpretation is crucial.

6.2 Limitations and Future Directions

This study's primary limitation stems from the monthly frequency of the U.S. Bureau of Labor Statistics (BLS) data, which fundamentally restricted the temporal granularity of our analysis. This misalignment likely obscured finer-grained causal relationships. Our findings also reflect the demographic and behavioral characteristics of Reddit users, who do not necessarily represent the broader labor force. Further, the struggle of the BERTopic model to cluster content in the Leisure & Hospitality sector indicates an example of how common NLP methods may be insufficient.

Future research should aim to incorporate labor market and unemployment data at higher temporal granularity (e.g., daily, weekly) to better align with social media activity. Additionally, researchers could utilize improved analytic approaches to better capture the precise direction and timing of the relationships. Expanding the scope of analysis to include broader economic signals (e.g., large-scale layoffs or financial market fluctuations) would further contextualize shifts in public discourse. Although the present study is exploratory, it establishes a useful foundation for fine-grained investigations at the intersection of social media and labor market research.

An extension to this work would be to implement a more fine-grained temporal segmentation to analyze social media features within distinct economic regimes (e.g. periods of expansions, volatility, or contraction within sectors), specifically with phase analysis. Such regime-aware analysis is significant because the relationship between online discourse and labor market indicators is unlikely to remain stable over time. Incorporating this approach could reveal more consistent correlations and potentially strengthen Granger causality results by identifying when predictive relationships fluctuate across different economic conditions.

7 Conclusion

In this paper, we present FLOUR: Fluctuations in Labor-market Outcomes & Unemployment Related-sentiment. We conduct an investigation into relationships between online sentiment and unemployment rate, seeking to understand discourse and sentiment surrounding economic trends across five sectors: Information, Education & Health, Leisure & Hospitality, Financial, and Agriculture. We collect a dataset of over 400,000 Reddit posts, running VADER sentiment, LIWC, autoregressive modeling, and BERTopic modeling. We run Pearson correlation and Granger causality tests on these results. Our findings are heavily sector-dependent. In Information

& Finance communities, we find significant correlations between work or money lexicon and unemployment. In Education & Health communities, we find a surprising inverse causal relationship between sentiment and unemployment (less negative sentiment when unemployment rates increase). This is confirmed by autoregressive modeling in Education & Health, indicating predictive potential using community sentiment. Topic modeling and manual selection of relevant topics reveals limited significant results, likely due to the granularity of our data. Overall, this work suggests that social media platform communities can be valuable indicators or predictors of sector-specific economic stability. However, our findings are limited by the granularity of monthly unemployment data. Future research could uncover more nuanced relationships by incorporating finer data or analyzing specific economic periods. By extending our data collection and analysis frameworks, future work can build more robust, sector-specific predictive models that best reflect emerging economic trends.

8 Contribution Statement

All group members participated equally in ideation of the project, presentation of the project during the poster session, and writing the Abstract, Introduction, and Conclusion.

Evan scripted Reddit data collection across all project stages and maintained the project's GitHub repository. He cleaned up final code notebooks and wrote the BERTopic model methodology for RQ2. For the final report, Evan wrote most of the Data and Methods sections, managed the bibliography, and handled formatting for Overleaf. Evan also provided project management by defining equal task distributions and ensuring alignment with project deadlines during weekly meetings.

Rachel scripted the Granger Causality and Pearson correlation analysis code, implemented the autoregressive models, and calculated SMAPE metrics. Additionally, she wrote several sections of the proposal, including the incorporation of feedback, single paragraph collaboration plan, related works, and intro. She also wrote the sections on autoregressive models that appear in the final paper. Finally, she contributed to designing the poster alongside Emily.

Emily outlined potential analysis methods in the initial stages of our project, ran sentiment and LIWC analyses, added basic plots for the significant correlation and causation sections identified, and wrote code to visualize and plot significant results. For the final report, Emily wrote the preliminary analysis, summary of findings, implications, and the limitations and future implications sections.

Ethan discovered and collected FRED unemployment data, defined project timeline for proposal, implemented preliminary sentiment and LIWC results sections, imported and formatted all Appendix figures to Overleaf. Ethan attempted to run causal impact package to evaluate results but the results were limited in significance and scope and thus results were omitted. For the final paper, Ethan contributed to the conclusion and methods: specifically, the statistical tests.

References

[1] 2025. Federal Reserve Economic Data (FRED) Release Tables. https://fred.stlouisfed.org/release/tables?rid=50&eid=4635&od=2025-06-01# Accessed October 2025.

[2] 2025. Industry at a Glance by NAICS Codes. https://www.bls.gov/iag/tgs/iag_index_naics.htm Accessed October 2025.

[3] 2025. Subreddit Stats Website. https://subredditstats.com/ Accessed October 2025.

[4] Hongyu An, Xiao Gu, Bojan Obrenovic, and Danijela Godinic. 2023. The Role of Job Insecurity, Social Media Exposure, and Job Stress in Predicting Anxiety Among White-Collar Employees. *Psychology Research and Behavior Management* 16 (2023), 3303–3318. doi:10.2147/PRBM.S416100

[5] Michelle Bak, Chungyi Chiu, and Jessie Chin. 2023. Mental Health Pandemic During the COVID-19 Outbreak: Social Media As a Window to Public Mental Health. *Cyberpsychology, Behavior, and Social Networking* 26, 5 (May 2023), 346–356. doi:10.1089/cyber.2022.0116 Epub 2023 Apr 13.

[6] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker. 2022. *The development and psychometric properties of LIWC-22.* Technical Report. University of Texas at Austin, Austin, TX. https://www.liwc.app

[7] C. W. J. Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* 37, 3 (1969), 424–438. doi:10.2307/1912791

[8] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[9] Arthur Heitmann. 2024. Arctic Shift API. https://github.com/ArthurHeitmann/arctic_shift/tree/master/api Accessed October 2025.

[10] Tong Liu, Christopher M. Homan, Cecilia Ovesdotter Alm, Ann Marie White, Megan C. Lytle, and Henry A. Kautz. 2016. Understanding Discourse on Work and Job-Related Well-Being in Public Social Media. In *Proceedings of the Conference of the Association for Computational Linguistics.* Association for Computational Linguistics, 1044–1053. doi:10.18653/v1/p16-1099

[11] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. Social Media Fingerprints of Unemployment. *PLOS ONE* 10, 5 (2015), e0128692. doi:10.1371/journal.pone.0128692

[12] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2, 11 (2017), 205. doi:10.21105/joss.00205

[13] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426* (2018). https://arxiv.org/abs/1802.03426

[14] Jamal Senouci. 2023. *causalimpact: A Python package for causal inference using Bayesian structural time-series models.* https://pypi.org/project/causalimpact/ Python port of the R package CausalImpact.

[15] Narendranath Sukhavasi, Janardan Misra, Vikrant Kaulgud, and Sanjay Podder. 2023. Geo-sentiment Trends Analysis of Tweets in Context of Economy and Employment during COVID-19. *Journal of Computational Social Science* (2023). doi:10.1007/s42001-023-00201-2

[16] Miaoqing Tan, Zhigang Wu, Jin Li, Yuxi Liang, and Wenting Lv. 2024. Analyzing the Impact of Unemployment on Mental Health among Chinese University Graduates: A Study of Emotional and Linguistic Patterns on Weibo. *Frontiers in Public Health* 12 (2024), 1337859. doi:10.3389/fpubh.2024.1337859

A GitHub Repository

All components of this project, including data (collection, aggregation, and cleaning), experimental methods, and final results are transparently documented and hosted in a public GitHub repository, accessible at: https://github.com/zuyouchen/cs598_cwb_final_proj.

A.1 Data Collection Script

To prevent rate-limit violations, posts were collected in batches of 100 using 0.25 second timeouts between calls. Each response was parsed into a structured format, including post text and metadata (e.g., author, score, upvotes, creation timestamp, permalink). Posts without textual content (e.g., link or image-only submissions) and entries with missing essential metadata (e.g. ID, creation time) were excluded. Our data collection script is available in our GitHub repository as `data_collection.ipynb`.

B Additional Figures

Additional figures not embedded in the paper above are included in their respective section of each analysis method. They can also be found in our GitHub repository in `/results`.

B.1 VADER Sentiment

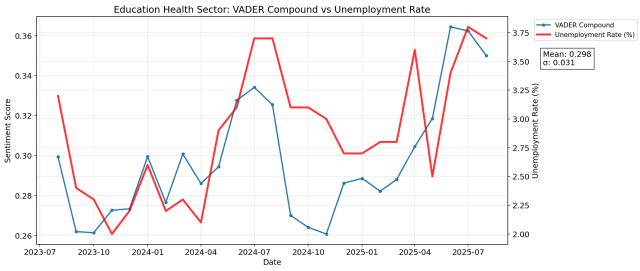

Figure 7: Education and Health - VADER Compound

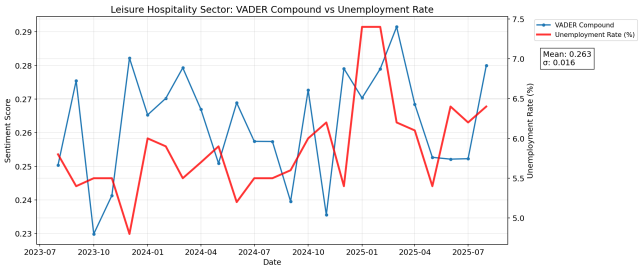

Figure 8: Leisure Hospitality - VADER Compound

B.2 LIWC

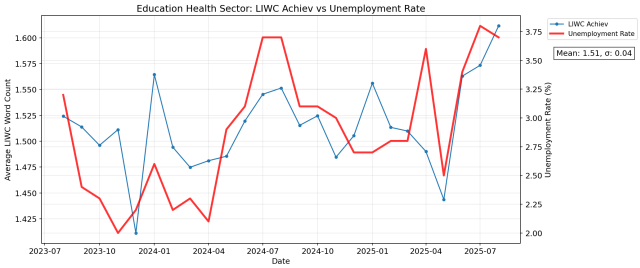

Figure 9: Education and Health - LIWC Achievement

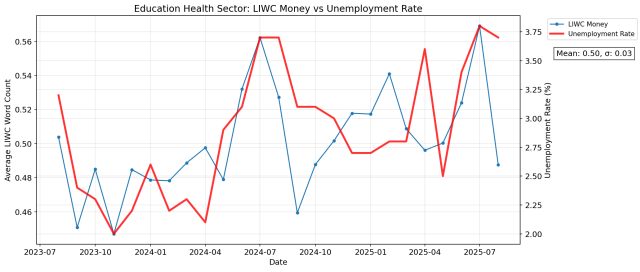

Figure 10: Education and Health - LIWC Money

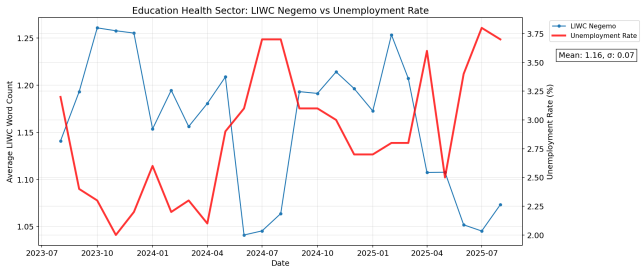

Figure 11: Education and Health - LIWC Negative Emotion

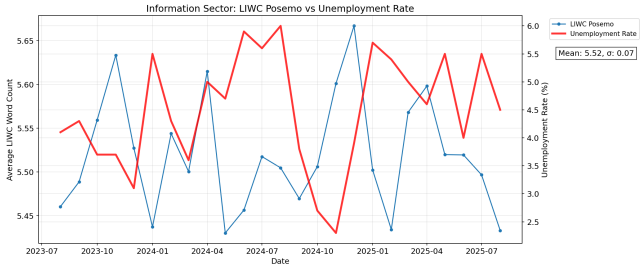

Figure 12: Information - LIWC Positive Emotion

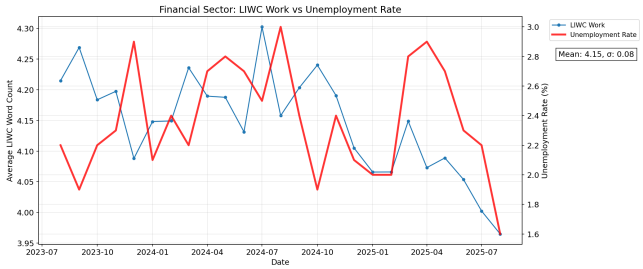

Figure 13: Financial - LIWC Work

B.3 Autoregressive Modeling

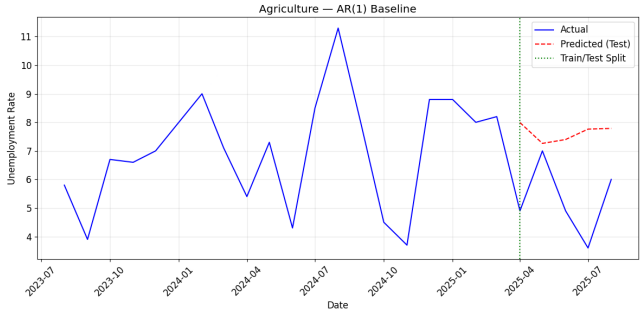

Figure 14: Agriculture AR - FRED + LIWC

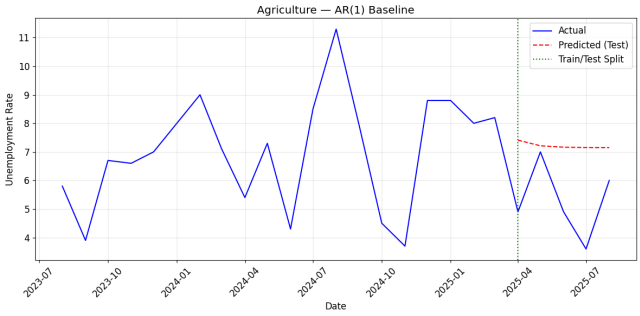

Figure 15: Agriculture AR - FRED Only

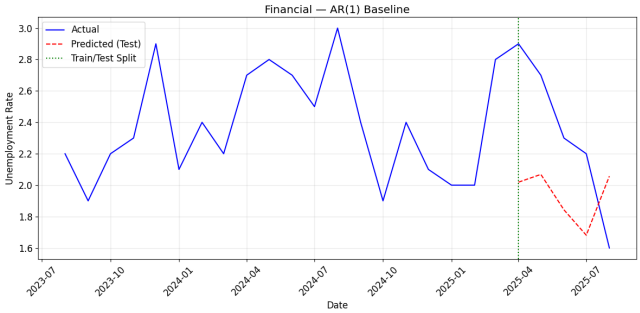

Figure 16: Financial AR - FRED + LIWC

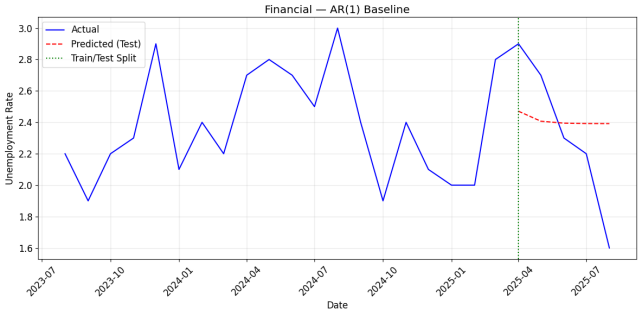

Figure 17: Financial AR - FRED Only

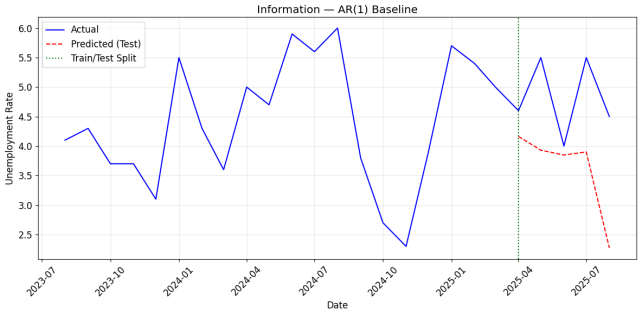

Figure 18: Information AR - FRED + LIWC

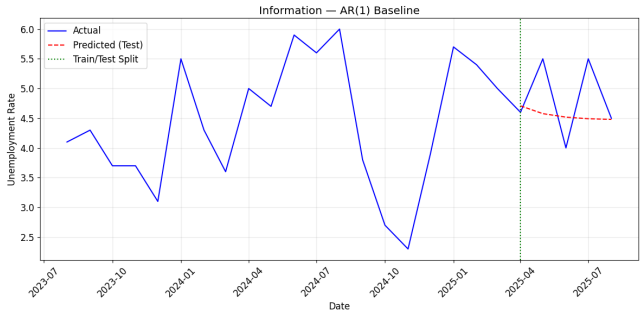

Figure 19: Information AR - FRED Only

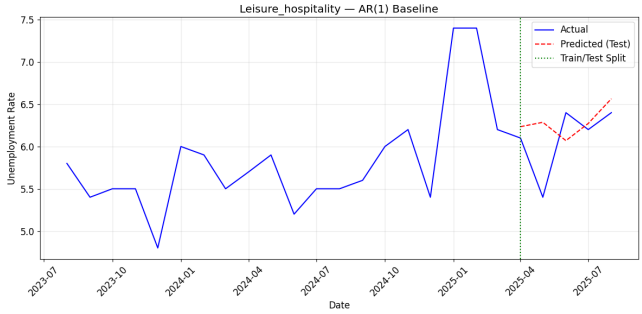

Figure 20: Leisure Hospitality AR - FRED + LIWC

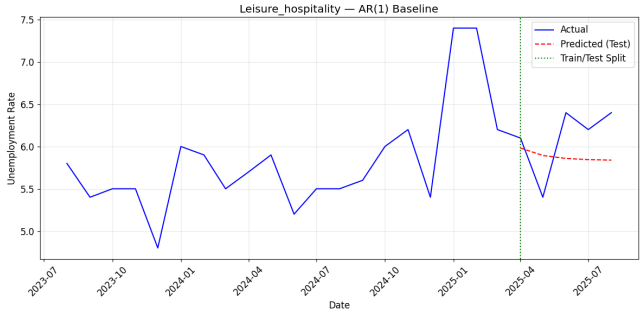

Figure 21: Leisure Hospitality AR - FRED Only

B.4 BERTopic

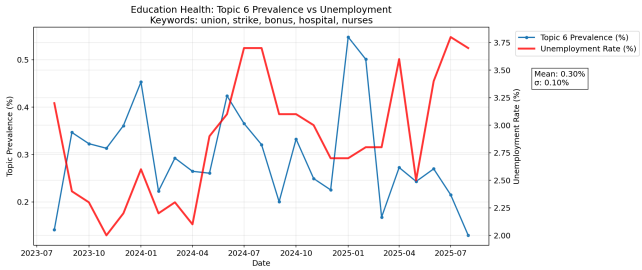

Figure 22: Education Health Topic 6 Prevalence

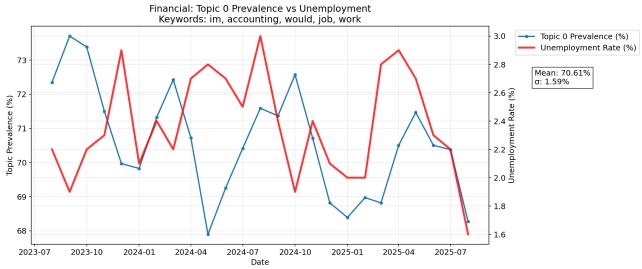

Figure 23: Financial Topic 0 Prevalence

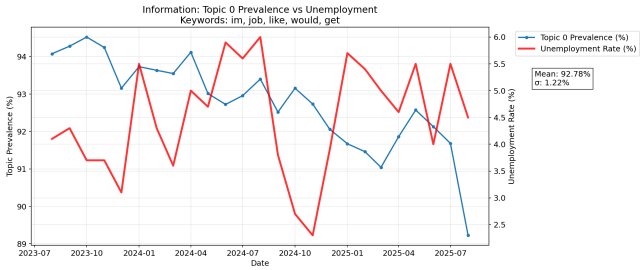

Figure 24: Information Topic 0 Prevalence

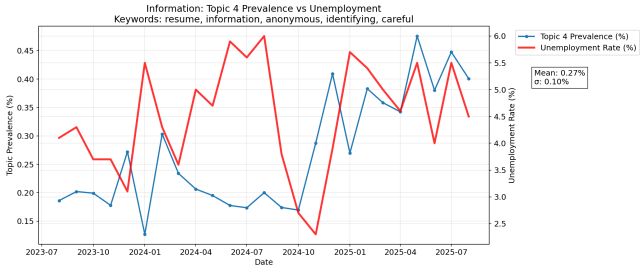

Figure 25: Information Topic 4 Prevalence