Covid and Overall Inmate Sentiment

Tim Huang, You Li, Xiang Li, Eileen Chang

December 11, 2023

The code for this project can be found at https://github.com/TheNOOONShy/DS-CapStone-Project-Group9.

Abstract

This research delves into the often-overlooked realm of the incarcerated population's experiences during the COVID-19 pandemic, shedding light on the unique challenges faced within correctional facilities. With a focus on both jails and prisons, our study explores the sentiments expressed in communications obtained through the PrisonPandemic[6] project, coupled with COVID-19 data from California. The analysis reveals distinct patterns in communication dynamics, with factors such as death rates and communication type playing pivotal roles in shaping sentiments. The XGBoost model further unravels the intricate interplay of these factors, emphasizing the nuanced impact within the incarcerated communities. As our findings deepen the understanding of the pandemic's repercussions on this marginalized group, we advocate for comprehensive research that amplifies the voices of the incarcerated in discussions surrounding public health and social well-being.

1 Introduction

In exploring the societal repercussions of the COVID-19 pandemic, one demographic often overlooked is the incarcerated population. While numerous studies have focused on the emotional and mental health effects on the general public, little attention has been given to those behind bars.

Correctional facilities present a unique challenge during a pandemic due to close quarters and limited healthcare resources. Overcrowding increases the risk of virus transmission, exacerbating the stress in an already tense environment. Despite these challenges, there is a noticeable absence of research on how the pandemic has affected the emotions of incarcerated individuals.

This research aims to fill that gap by investigating the specific impact of the COVID-19 pandemic on the emotions of those in prison. We delve into this neglected aspect to understand the unique challenges faced by the incarcerated during the pandemic, with the goal of providing insights that can inform improved support systems and policies for this marginalized group. As we navigate the enduring consequences of the pandemic, it is crucial to consider the experiences of those within prison walls, ensuring their voices are heard in discussions about public health and social well-being.

2 Data Description

2.1 Data Sources

For our study, we needed to gather data from multiple sources. The primary data, which includes the texts for evaluating prisoner sentiments, was obtained from the PrisonPandemic Project[6]. It's important to note that none of the transcripts included COVID data. Consequently, we had to source the COVID-related data from different databases.

2.1.1 PrisonPandemic

From the PrisonPandemic dataset, we obtained a total of 771 files from both jails and prisons. The majority of the data, comprising 642 communications, originated from prisons, while the remaining 129 came from jails. Notably, an intriguing detail emerged from the received jail communications, revealing only 4 positive sentiments out of the entire 129. This suggests that less than 4% of sentiments in jail communications were positive (refer to Table 1). Additionally, we conducted an analysis of the distribution between calls and letters, and the exact numbers can be found in Table 2.

	Prison	Jail	Total
Negative	582	125	707
Positive	60	4	64

Table 1: Sentiment Distribution

	Call	Letter
Count	553	219

Table 2: Call and Letter Counts

2.1.2 COVID data

We obtained COVID data, including infection and death statistics, from the California Department of Corrections and Rehabilitation (CDCR) for prisons[4]. However, tracking COVID cases in jails, where individuals frequently enter and exit, poses challenges. Consequently, for jail-related COVID data, we relied on county-level information obtained from USA Facts[8].

3 Data Processing

3.1 Stories to Sentiments

Before our in-depth analysis, we transformed PrisonPandemic narratives into sentiment values for model processing. Despite exploring various models, including NLTK, challenges arose, with NLTK's Vader Score often mislabeling sentiments as "neutral" and showing a bias toward classifying more positive sentiments.
After thorough exploration, we identified Flair[1] as the most promising model. Given the dual nature of our dataset (interviews and letters), distinct processing methods were necessary. For calls, where UCI members posed questions to incarcerated individuals, we filtered to include only prisoner statements to mitigate potential sentiment bias. For letters, entirely authored by inmates, we retained all information, making minimal adjustments for Flair's sentiment analysis.

3.2 Covid-19 Data Combined with PrisonPandemic data

There were some issues with the date column in the data from PrisonPandemic, and we also aimed to accurately capture the correct prison/jail names. To address this, we extracted the date and prison/jail names from the title of the data in PrisonPandemic. However, since the prison names in both data sources had different formats and the PrisonPandemic data itself exhibited varied formats, we needed to standardize the name formats in PrisonPandemic to ensure uniform patterns. Additionally, we defined the weeks starting from January 1st, 2020 (i.e., 1/1/2020 - 1/7/2020 would be one week, and 1/8/2020 - 1/15/2020 would be the next). This allowed us to clean up the name patterns in the COVID data and then match these two datasets based on prison/jail names and the defined weeks, generating a new dataset upon which we would be able to conduct our analysis.

3.3 Determining Peak COVID-19 Weeks in California

Identifying peak periods of COVID-19 was crucial, as these periods of heightened viral spread might correlate with significant changes in inmate behavior or mental health.
Given the nature of our data, with many weeks reporting zero new cases, calculating the mean of weekly new cases could be significantly skewed. This could lead to an unreliable baseline for peak determination due to distortions from weeks with no reported cases or unusually high case numbers (outliers). To address it, this is the process of how we identify the peak weeks:

1. **Median Calculation**: Computing the median of the weekly new cases to establish a central trend line less affected by zero-case weeks and outliers.
2. **Standard Deviation Computation**: Calculating the standard deviation of the weekly new cases to measure the variability around the median.
3. **Threshold Establishment**: Defining the peak week threshold as the sum of the median and the standard deviation, marking weeks surpassing this sum as peak periods of COVID-19 spread.
4. **Peak Week Identification**: Marking weeks with new case numbers exceeding the threshold as peak weeks in the `is_Covid_Peak` column.

This methodology enables us to more accurately identify when the pandemic had a significant impact, a critical aspect for analyzing its effects on the inmate population (See Figure 1).

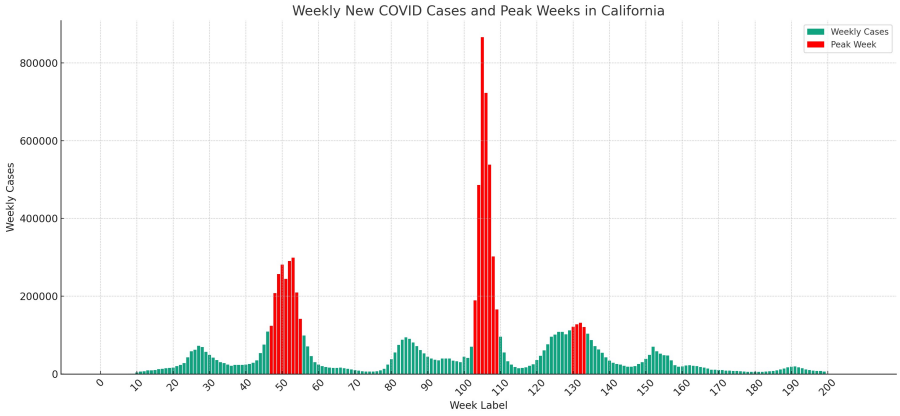


Figure 1: Peak Week

3.4 Handling Imbalance

Upon completing the data processing, a notable imbalance[7] emerged in the number of positive and negative sentiments, with a ratio of almost 10:1. Addressing this imbalance before modeling was crucial to enhance the robustness of our models. Given our limited dataset of under 200 samples for the minority class and the need to create descriptive models, undersampling the majority was deemed impractical. Consequently, we opted for oversampling.

3.4.1 Individual Sentiment Prediction

One of the models we aimed to develop was a logistic regression model [3] to determine the log-odds of factors predicting positive or negative sentiments. Recognizing the sensitivity of logistic regressions to data imbalance, we chose to handle the imbalance by oversampling positive sentiment texts. As our samples already reflected the type of COVID conditions corresponding to the communication we ended up deciding to go with simple random oversmapling. However given that there were very few samples for jails, SMOTE or ADASYN would have been reasonable methods of oversampling as well. As such, we will provide non-oversampled results as well as the oversampled results for jail results.
For the random oversampling, we conducted it with a 1:1 ratio. We also created models without oversampling, but these models exhibited significantly worse metrics. The extent of oversampling varied between prisons and jails due to the ratios being different.

3.4.2 Group Sentiment Prediction

Another model we attempted to create was a linear model to assess the correlation between different factors and the number of correspondences received by PrisonPandemic. Three distinct models were developed to evaluate the number of positive sentiment correspondences, negative correspondences, and total correspondences. Given the focus on the quantity of correspondences, no oversampling or undersampling of the data was deemed necessary.

4 Statistical Methods Utilized

We explored various statistical methods to identify potential relationships between COVID and prisoner sentiment. Due to the limited availability of key features, the models we employed were relatively straightforward.

4.1 Linear Regression

One primary concern in our data analysis was the potential presence of multicollinearity among factors such as case rate, death rate, and population. While some correlation between these factors is expected, our examination revealed that, for jails, there was moderate multicollinearity. However, the Variance Inflation Factor (VIF) scores were not excessively high, allowing us to derive conclusive results (refer to Table 3). Conversely, prisons exhibited significantly high VIF scores when considering both the number of cases and deaths. To address this issue, we opted to exclude the number of cases, resulting in lower VIF scores (refer to Table 4).

Variable	VIF
Week Number	1.195791
County Cases	7.434936
County Deaths	6.870570
County Case Rate	2.348540
County Death Rate	1.326963

Table 3: Jail

Variable	VIF
Is State Prison	2.780516
Week Number	1.170324
Weekly Death Number	2.706860
Weekly Death Rate	1.110951
Weekly Case Rate	1.153018

Table 4: Prison

Table 5: Variance Inflation Factors (Linear Regression)

4.2 Logistic regression

Concerns regarding multicollinearity were relevant to both our logistic and linear regressions. However, examination of VIF scores for jails (Table 7) and prisons (Table 6) revealed values consistently below 5, with only two factors breaching six, signifying relatively low multicollinearity. The disparity in VIF values between the linear and logistic models may stem from our linear regression's facility-level approach, contrasting with the logistic regression conducted at the individual communication level. Despite both using a weekly cadence of cases and deaths, assessing communications individually introduced finer granularity in logistic regression, potentially mitigating multicollinearity. VIF values for both jail and prison logistic regressions align with accepted thresholds, significantly below 10, indicating acceptable levels. Hence, despite differences in data nature and regression models, our analysis suggests minimal multicollinearity impact on logistic regression stability, bolstering confidence in our results.

Variable	VIF
Number of Cases	6.227430
Case Rate	2.316808
Number of Deaths	6.839367
Death Rate	1.256819
File Lengths (Char)	1.050545
Week Number	1.337324
VADER Score [5]	1.103226
Is Covid Peak	1.525807
Prison Max Size	2.245943
Is Letter	1.097596

Table 6: Prison

Variable	VIF
Number of Cases	3.108155
Number of Deaths	4.169468
File Lengths (Chars)	1.190993
Week Number	2.413715
Vader Score	1.494336
Is COVID Peak	1.858614
County Population	3.094234
Is Letter	2.678458

Table 7: Jail

Table 8: Variance Inflation Factors (Logistic Regression)

4.3 XGBoost

In our final analysis, our team employed an XGBoost[2] model, recognized for its predictive accuracy, to assess the significance of individual factors within our dataset. Shifting the focus from facility-level perspectives to individual-level dynamics, this approach aimed to identify key predictors and measure their impact on overall predictive performance. By examining feature importances, we gained insights into intricate relationships and nonlinear patterns present in our data, specifically regarding individual-level interactions. This strategic shift in perspective enhanced our understanding of the complex dynamics governing the relationship between COVID and prisoner sentiment, contributing to the depth and precision of our analytical approach.

5 Results

5.1 Linear Regression (Individual)

There was an attempt made to use the confidence of Flair as a measure of how positive or negative a sentiment any given communication received was. However, we found that the confidence was not a good indicator (upon reading the texts), and many communications had the same values for their factors. Consequently, the models created resulted in very low R-squared values (under 0.01). As such, we pivoted away from trying to gauge an

individual's sentiment strength and simply created new models treating sentiment as either just positive or just negative.

5.2 Linear Regression (Communication Dynaimcs)

5.2.1 Jails

One of the key issues when it came to running a weekly linear regression for the number of communications received from jails was the limited number of positive communications (only 4). Consequently, the regression for positive counts resulted in no statistically significant features (see Figure 4) as well as minimal impact on the total weekly counts (compare Figure 2 and Figure 3). However, even for the two models that did result in statistically significant factors, there is still a lot of correlation that was not accounted for by the factors we were able to generate.

Because the total counts and weekly counts are so similar, the analysis for jail counts is mainly focused on the negative counts. While most of the factors were statistically significant at the 0.1 p-value level, we found that the weekly county death rate really did not have any statistical impact on the number of communications received from a given jail in a given week. In addition, it looks as if the number of cases and the number of deaths were more statistically important in the number of communications received per week. This may be due to the fact that a lot of COVID reporting is reported in terms of total cases/deaths rather than cases/deaths per 1000, and such reporting results in people wanting more contact with the outside when they sense that there is a large number of cases/deaths rather than a large rate of cases and deaths. One of the other items that can be seen is that the week number is statistically significant, with every week there being almost 1/2 a communication less from any given jail. This could be due to the fact that people started to get used to the COVID environment and do not feel the need to have to send communications out as much as when COVID first started.

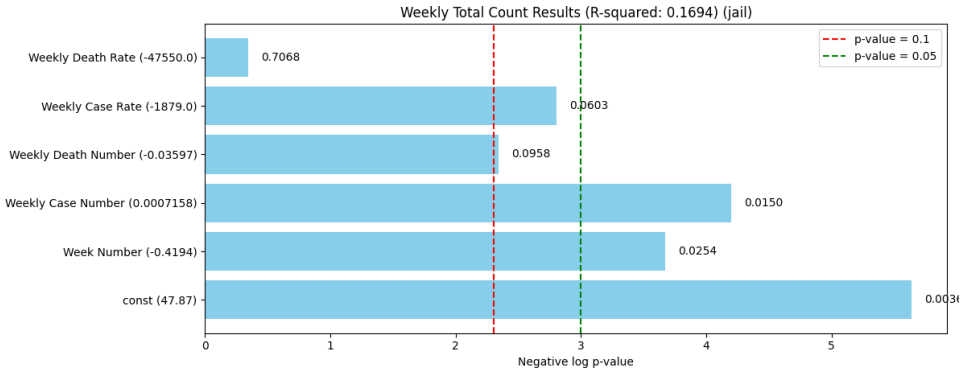


Figure 2: Weekly Counts Regression (Jail) (Table 11)

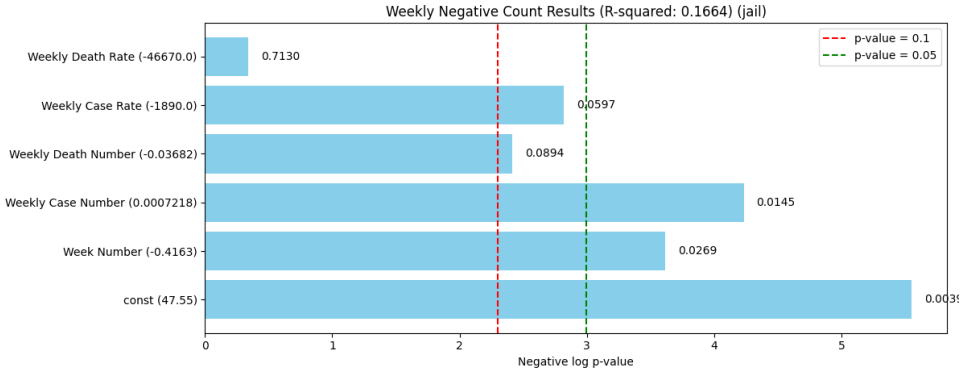


Figure 3: Weekly Negative Counts Regression (Jail) (Table 12)

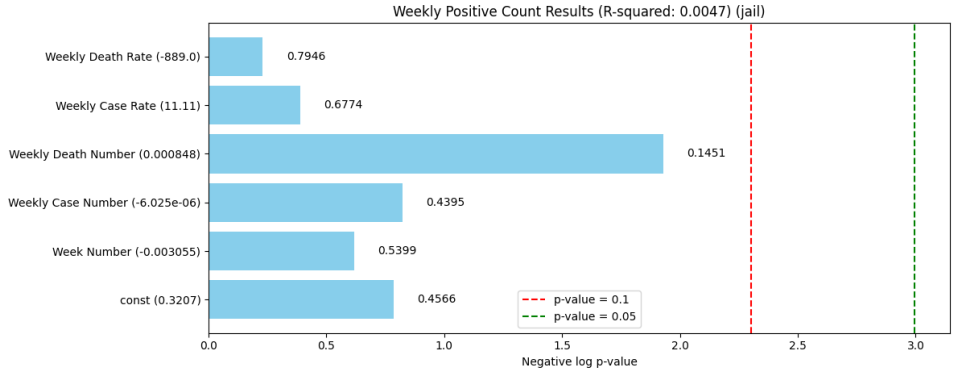


Figure 4: Weekly Positive Counts Regression (Jail) (Table 13)

5.2.2 Prisons

Regarding prisons, a sufficient number of negative and positive samples allowed the creation of relatively robust models for all three counts (refer to Figure 5, 6, and 7). Apart from identifying statistically significant factors, the factors used also contributed to explaining more correlation in the data.
A notable trend observed in all three models is that the number of deaths consistently emerged as more significant predictors for communication counts compared to death rates and case rates. An increase in deaths correlated with a rise in communications, potentially indicating that inmates are more inclined to share their experiences during times of increased mortality. We could not do the same analysis on the results of the number of cases, as adding in a number of cases results in very large VIF scores for both death number and case number.
An intriguing factor that drew our attention is the "Is State Prison" variable. Given the categorization of prisons into federal and state types, we hypothesized that the facility type might impact communication counts. Our findings indicate that, in a given week, a state prison is expected to send out more than 3 extra communications, with the majority of those being negative. While state prisons exhibited an increased number of negative sentiments, a comparable impact on positive sentiments was not observed. This suggests a potential trend where prisoners in state prisons tend to express more negativity for some reason.

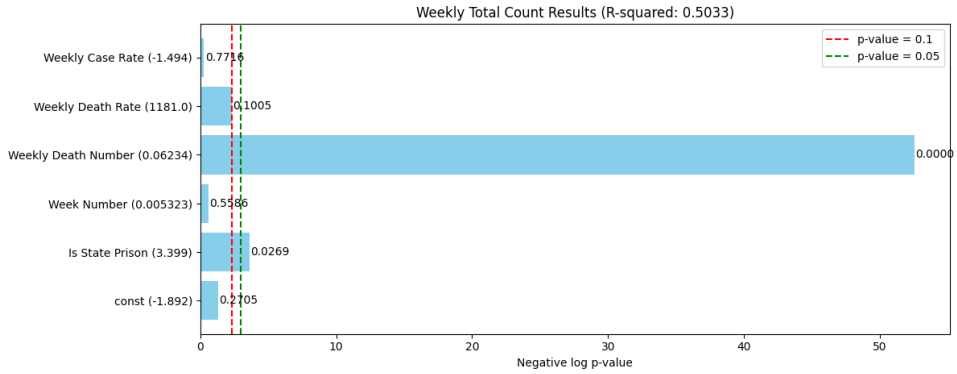


Figure 5: Weekly Counts Regression (Prison) (Table 14)

5.3 Logistic Regression (Individual)

5.3.1 Jail

The Jail regressions are divided into two sections due to a notable discrepancy in the number of positive and negative sentiments (See Figure 8a for oversampled regression and Figure 8b for non-oversampled results). Upon oversampling, the relative importance of each feature remains consistent, although individual p-values undergo changes, and log odds show slight variations. As the oversampled model enhances explainability and addresses logistic regression's issues, the subsequent analysis focuses on oversampled results (Figure 8a).

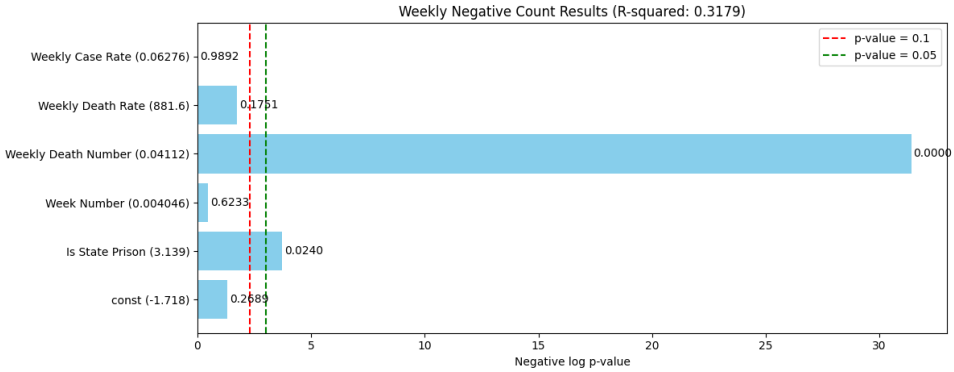


Figure 6: Weekly Negative Counts Regression (Prison) (Table 15)

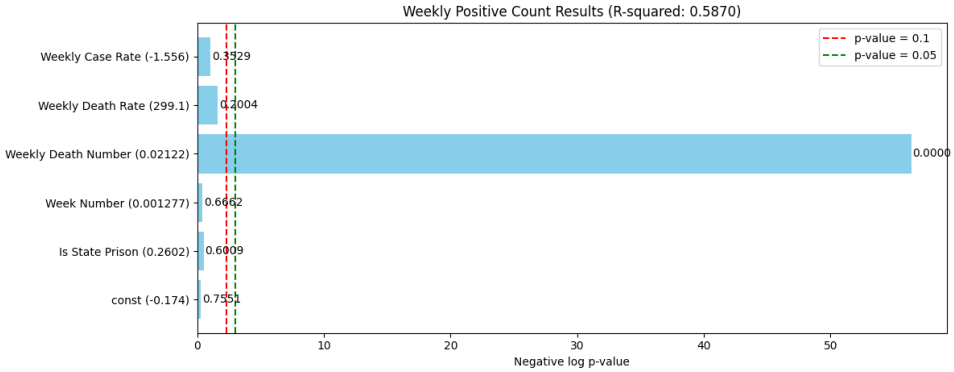


Figure 7: Weekly Positive Counts Regression (Prison) (Table 16)

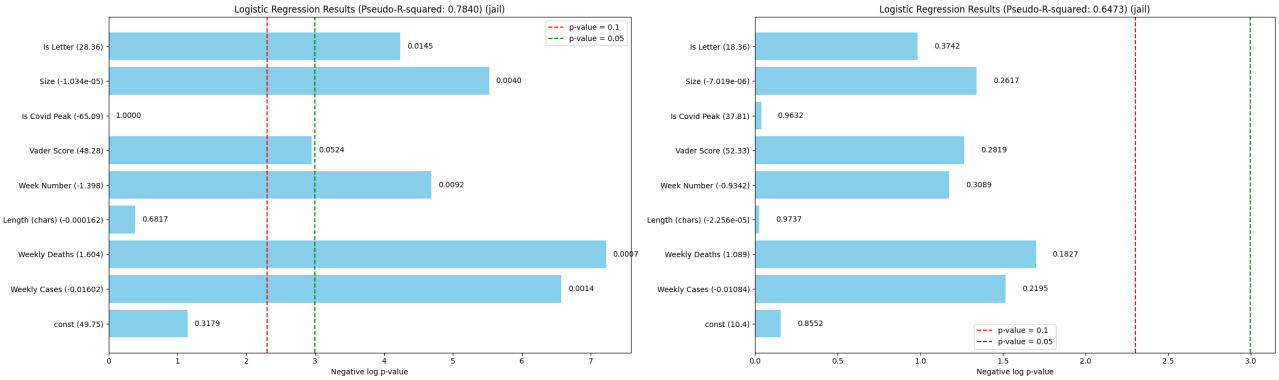
The regression reveals that almost all included features hold statistical significance. Examining the log-odds directions becomes crucial, where a positive log-odds value suggests an increased likelihood of positive communication, and vice versa for negative log-odds. In the context of assessing how COVID affects sentiments in jails, the analysis indicates that an increase in deaths leads to a higher likelihood of positive sentiments, while a decrease in the number of cases is associated with a higher likelihood of positive sentiments. The seemingly contradictory signals between deaths and cases may be explained by perceiving county-wide deaths as inevitable, resulting in a more positive sentiment towards the county's handling of COVID deaths, while an increase in cases may contribute to a more negative view as it is seen as something preventable by public officials.

5.3.2 Prison

For prisons, a single figure (Figure 9) suffices, as the disparity in positive and negative communications is less pronounced. Almost all features used are statistically significant, although the fit is not as robust as observed in jails. Focusing on COVID cases and deaths within prisons reveals a sign discrepancy between weekly case rate and number, as well as weekly death rate and number.
The negative sign of weekly deaths (and positive sign of weekly death rate) implies that a death within a smaller prison population has less negative impact on inmate sentiment compared to a larger population, suggesting that, in smaller populations, a death may even increase the likelihood of a positive sentiment. Conversely, the negative sign of weekly case rate (and positive sign of weekly cases) suggests that a single case has a more negative impact on inmates in smaller prisons, possibly due to the perception that deaths are less preventable than cases, or smaller facilities find it easier to stay positive with concrete outcomes (death) than unclear outcomes (cases).

5.3.3 Comparisons

When comparing the two models, it's notable that weekly rates are included for prisons but not for counties. This decision is driven by the slight difference in the "Size" factor between jails and prisons. For jails, size corresponds to the county's size, resulting in collinearity issues, while for prisons, size represents the maximum



(a) Oversampled (Table 17)

(b) No Oversampling (Table 18)

Figure 8: Logistic Regression Result Jail

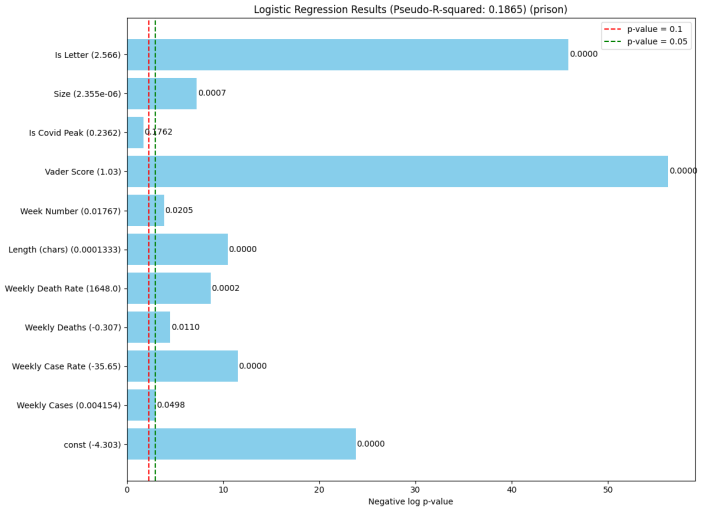


Figure 9: Logistic Regression Result Prison (Table 19)

capacity rather than the current number of inmates. In terms of log-odds sign similarities and differences, communications in the form of letters consistently yield a higher likelihood of positive sentiment in both jail and prison contexts. The VADER score, a sentiment score based solely on communication content, is positively correlated as expected, though it is less statistically significant for those in jail compared to those in prison. Notably, there is a sign flip for the log-odds of Size, Week Number, and Length, possibly attributed to the different measurement approaches for Size in counties versus prisons.

5.4 XGBoost

When we attempted to use the more complex XGBoost model to classify individual sentiments, we found that with our given factors, XGBoost was too complex and suffered from overfitting, as seen from the 100% accuracy in Table 9 and Table 9. However, even though there is overfitting, the feature importances can still provide us with information that we would be unable to gain from viewing a simple linear or logistic regression.

	Precision	Recall	F1-Score	Support
Negative	1.00	1.00	1.00	582
Positive	1.00	1.00	1.00	60
Accuracy			1.00	
Macro Avg	1.00	1.00	1.00	642
Weighted Avg	1.00	1.00	1.00	642

Table 9: Classification Metrics for XGBoost Prison

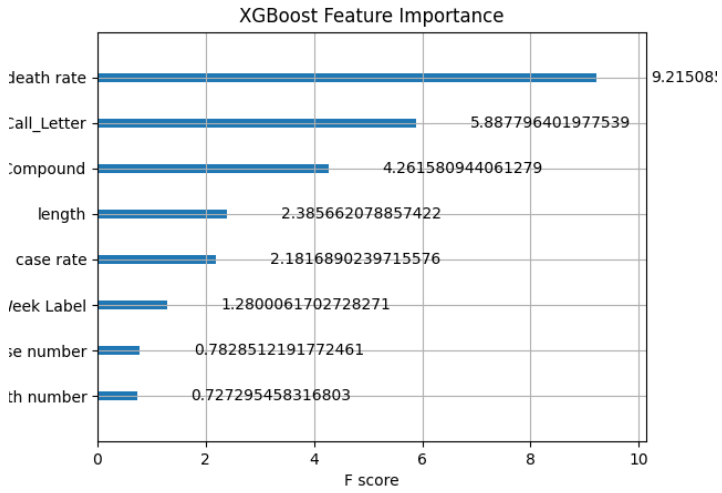


Figure 10: Gain of XGBoost Features (Jail) (See Table 20 for raw data)

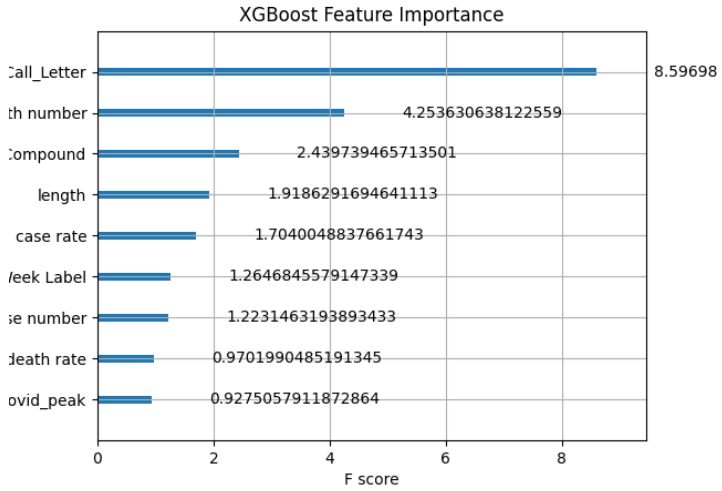


Figure 11: Gain of XGBoost Features (Prison) (See Table 21 for raw data)

5.4.1 Jail

We find that for jails, the most indicative features for whether a communication would be positive or negative were actually the death rate for the county and whether the communication is a call or a letter (See Figure 10). In a more complex model, death/case rates are more important than death/case numbers, which is what we originally suspected; however, the relationship does not appear to be linear. This indicates that there may be more research opportunities available in this direction: analyzing the relationship between the rates and jail inmate sentiments.

5.4.2 Prison

When we look at prisons (Figure 11), we find that it is a different story, and the most important features in determining if an inmate's story had a positive or negative sentiment were actually whether it was a call or a letter, followed by the number of deaths. While the feature importance does not tell us if the feature indicates more positive or negative sentiment, by pairing it with our logistic regression, we can reasonably say that a communication being a letter is a strong indication that the sentiment is more positive than a communication that was a call. It is interesting that within prisons, the death number is so much more important than the case number, case rate, or even death rate. This could indicate that the small community within a prison allows for

	Precision	Recall	F1-Score	Support
Negative	1.00	1.00	1.00	125
Positive	1.00	1.00	1.00	4
Accuracy			1.00	
Macro Avg	1.00	1.00	1.00	129
Weighted Avg	1.00	1.00	1.00	129

Table 10: Classification Metrics for XGBoost Jail

prisoners to relate to deaths more closely than someone in a jail, but more research into this would be needed.

6 Conclusion

Our study on the repercussions of COVID-19 on incarcerated populations yields significant insights:

6.1 Communication Dynamics

Jails: Weekly communication counts in jails were influenced by county cases, deaths, and the passage of time. Higher county death rates correlated with more positive sentiments.
Prisons: Communication counts in prisons were notably influenced by internal COVID-19 dynamics. State prisons exhibited a tendency to have more negative sentiments without any corresponding increase or decrease in positive sentiments compared to federal prisons.

6.2 Factors Shaping Sentiments

Jails: In jails, the number of county deaths was linked to an increase in positive sentiments, while a lower number of cases corresponded to more positive expressions. Additionally, letters were more likely to convey positive sentiments.
Prisons: Prison sentiments were influenced by an increase in internal deaths, indicating a unique perspective within the prison community. The communication type, especially letters, played a pivotal role in conveying positive sentiments.

6.3 XGBoost Model Insight

Our XGBoost model underscored the intricate interplay of factors. For jails, death rates and communication type emerged as crucial, while in prisons, communication type and death counts held key significance.

6.4 Challenges and Future Directions

Addressing imbalances in sentiment counts and navigating potential overfitting presented challenges. Future research avenues could explore refined sentiment analysis techniques and delve deeper into the intricate dynamics of prison communities. In summary, our study enriches our understanding of how incarcerated individuals navigate the complexities of the COVID-19 pandemic, emphasizing the importance of inclusive research in shaping conversations around public health and well-being.

Acknowledgments

- **PrisonPandemic (Christine/Naiomi):** Thank you for providing the communications from prisoners.
- **State of California and CDCR:** For access to the public COVID-19 data that facilitated our analysis.
- **Xiang and Eileen:** Thank you for your invaluable contributions to the data gathering and preprocessing.
- **You (our team lead):** Thank you for keeping us on track and identifying possible important features.
- **Tim:** Thank you for your efforts in model development, result analysis, and writing, editing, and formatting the final report.
- **Professor Chen Li and Pooya Khosravi:** Thank you for helping us develop our skills.

References

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Flair: A very fast state-of-the-art NLP library. *arXiv preprint arXiv:1808.08786*, 2018.

[2] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable and Accurate Implementation of Gradient Boosting. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[3] David R. Cox. Logistic regression: A statistical model for binary classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):358–375, 1979.

[4] Data.gov. Cdcr population covid-19 tracking. `https://catalog.data.gov/dataset/cdcr-population-covid-19-tracking/`, 2023.

[5] Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, 2014.

[6] Lattimore and Friends. Prison pandemic. `https://prisonpandemic.uci.edu/`, 2023.

[7] Guillaume Lemaître, Fábio Nogueira, and Chris K. Aridas. imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[8] OpenGov. Covid-19 time-series metrics by county and state. `https://data.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state-archived`, 2023.

7 Appendices

7.1 Raw Data Tables

Factor	Coefficient	P-value	0.1	0.05
Weekly Death Rate	-47550	0.7068		
Weekly Case Rate	-1879	0.0603	*	
Weekly Death Number	-0.03597	0.0958	*	
Weekly Case Number	0.0007158	0.015	*	*
Week Number	-0.4194	0.0254	*	*
const	47.87	0.0036	*	
R-squared				0.1694

Table 11: Weekly Total Count Regression Result (Jail)

Factor	Coefficient	P-value	0.1	0.05
Weekly Death Rate	-46670	0.713		
Weekly Case Rate	-1890	0.0597	*	
Weekly Death Number	-0.03582	0.0894	*	
Weekly Case Number	0.0007218	0.0145	*	*
Week Number	-0.4163	0.0269	*	*
const	47.55	0.0039	*	
R-squared				0.1664

Table 12: Weekly Negative Count Regression Result (Jail)

Factor	Coefficient	P-value	0.1	0.05
Weekly Death Rate	-889	0.7946		
Weekly Case Rate	11.11	0.6774		
Weekly Death Number	0.000848	0.1451		
Weekly Case Number	-6.03E-06	0.4395		
Week Number	-0.003055	0.5399		
const	0.3207	0.4566		
R-squared				0.0047

Table 13: Weekly Positive Count Regression Result (Jail)

Feature	**Coefficient**	**P-value**	**Significance**
Is State Prison	3.3995	0.027	*
Week Number	0.0053	0.559	
Weekly Death Number	0.0623	0.000	**
Weekly Death Rate	1180.7440	0.101	
Weekly Case Rate	-1.4937	0.772	

*Note: * Significant at 0.1 level, ** Significant at 0.05 level*

Adjusted R-squared: 0.503

Table 14: Weekly Total Count Regression Result (Prison)

Feature	**Coefficient**	**P-value**	**Significance**
Is State Prison	3.1393	0.024	*
Week Number	0.0040	0.623	
Weekly Death Number	0.0411	0.000	**
Weekly Death Rate	881.6287	0.175	
Weekly Case Rate	0.0628	0.989	

*Note: * Significant at 0.1 level, ** Significant at 0.05 level*

Adjusted R-squared: 0.318

Table 15: Weekly Negative Count Regression Result (Prison)

Feature	**Coefficient**	**P-value**	**Significance**
Is State Prison	0.2602	0.601	
Week Number	0.0013	0.666	
Weekly Death Number	0.0212	0.000	**
Weekly Death Rate	299.1153	0.200	
Weekly Case Rate	-1.5564	0.353	

*Note: * Significant at 0.1 level, ** Significant at 0.05 level*

Adjusted R-squared: 0.587

Table 16: Weekly Positive Count Regression Result (Prison)

Factor	Coefficient	P-value	0.1	0.05
Is Letter	28.36	0.0145	*	*
Size	-1.03E-05	0.004	*	*
Is Covid Peak	-65.09	1		
Vader Score	48.28	0.0524	*	
Week Number	-1.40	0.0092	*	*
Length (chars)	-0.000162	0.6817		
Weekly Deaths	1.604	0.0007	*	*
Weekly Cases	-0.01602	0.0014	*	*
const	49.75	0.3179		
Pseudo R-squared				0.7840

Table 17: Logistic Regression Results (Jail, Oversampled)

Factor	Coefficient	P-value	0.1	0.05
Is Letter	18.36	0.3742		
Size	-7.02E-06	0.2614		
Is Covid Peak	37.81	0.9632		
Vader Score	52.33	0.2819		
Week Number	-0.93	0.3089		
Length (chars)	-2.26E-05	0.9737		
Weekly Deaths	1.089	0.1827		
Weekly Cases	-0.01084	0.2195		
const	10.4	0.8552		
Pseudo R-squared				0.6473

Table 18: Logistic Regression Results (Jail, Not Oversampled)

Factor	Coefficient	P-value	0.1	0.05
Is Letter	2.566	0	*	*
Size	2.36E-06	0.0007	*	*
Is Covid Peak	0.2362	0.1762		
Vader Score	1.03	0	*	*
Week Number	0.01767	0.0205	*	*
Length (chars)	1.33E-04	0	*	*
Weekly Death Rate	-0.307	0.0002	*	*
Weekly Deaths	-0.307	0.011	*	
Weekly Case Rate	-35.65	0	*	*
Weekly Cases	0.004154	0.0498	*	
const	-4.303	0	*	*
Pseudo R-squared				0.1865

Table 19: Logistic Regression Results (Prison, Oversampled)

Feature Name	**Readable Feature Name**	**Feature Importance (Gain)**
death rate	Weekly Death Rate	9.215
Call_letter	Is Call/Letter	5.888
Compund	VADER Score	4.262
length	Length (Chars)	2.386
case rate	Weekly Case Rate	2.182
Week Label	Week Number	1.28
case number	Weekly Case Number	0.783
death number	Weekly Death Number	0.727

Table 20: Jail XGBoost Importance

Feature Name	**Readable Feature Name**	**Feature Importance (Gain)**
Call_letter	Is Call/Letter	8.597
death number	Weekly Death Number	4.254
Compund	VADER Score	2.44
length	Length (Chars)	1.919
case rate	Weekly Case Rate	1.704
Week Label	Week Number	1.265
case number	Weekly Case Number	1.223
death rate	Weekly Death Rate	0.97
covid_peak	Is Peak Covid	0.928

Table 21: Prison XGBoost Importance

Feature	**Meaning**
Weekly Death Rate	Deaths per population for the week
Weekly Case Rate	Cases per population for the week
Is Letter	1 if communication is a letter, 0 if it is a call
Week Number	Number of weeks since the PrisonPandemic project first received communication
Length (Chars)	Number of characters in communication
Weekly Case Number	Number of cases within the week
Weekly Death Number	Number of deaths within the week
Is Peak Covid	Refer to Section 3.3
Size (Jail)	County Size
Size (Prison)	Maximum prison capacity
Vader Score	Sentiment analysis score using the Vader sentiment analysis tool.
Is State Prison	1 if state prison, 0 if federal

Table 22: Features and Meanings