Leveraging Social Media Data for Public Health: A Study in Forecasting Opioid-Related Deaths

Noah Kader
April 25,2023
GitHub: https://github.com/nkader12/DAProject2023
CSCI 4440-02

Abstract

The opioid crisis has become a major public health concern worldwide, with thousands of people losing their lives every year due to opioid overdose. Despite various efforts to curb the crisis, the number of opioid-related deaths continues to rise. As such, it is essential to explore new and innovative ways of addressing this issue. The use of social media data as a source of information is a promising alternative that could lead to new insights and strategies for combating the opioid crisis. For more information on prior work in using social media as a sensor for real world phenomena refer to the Appendix A.

In this study, we investigate the potential of using Reddit Opiate forums as a sensor to predict opioid-related deaths. Our study involved the collection and analysis of data from these forums, including text data, time and date of posts, user activity, and sentiment analysis. We utilized a range of machine learning techniques, such as classification, multivariate regression, and clustering to gain insights from the data. Our trained classification models achieved a reasonable level of accuracy, while the regression models showed less success in predicting the number of opioid-related deaths. Additionally, topic clustering proved useful in grouping text data into meaningful topics.

Our findings suggest that social media data has significant potential for forecasting phenomena and can provide valuable information for predicting and preventing opioid-related deaths. Future research in this field is crucial to fully harness the possibilities of this data.

Data Description and Exploratory Data Analytics

To evaluate social media as a sensor for opioid-related discourse, we chose to analyze the subreddit forum Opiates.[5] The forum's anonymous nature made it an ideal source of data for our study, and the ease with which Reddit forums can be crawled facilitated data collection. We focused our analysis on the comments data frame, which contains 3,845,116 rows and 10 features, with the body, year, and month being the primary focus.

User	Body	C	Score	E	Timestamp	Year	Month
[deleted]	at bonnaroo. it smelled like flowers but I did...	1308881233	1	opiates	2011-06-23 22:07:13	2011	6
nvrwastetree	Lol methadone....abiet old school??	1308877576	1	opiates	2011-06-23 21:06:16	2011	6
ttd	Immodium (loperamide) does in fact have activi...	1308861650	1	opiates	2011-06-23 16:40:50	2011	6
nvrwastetree	Heard of this as well...	1308854583	1	opiates	2011-06-23 14:43:03	2011	6
enthreeoh	You're getting placebo effect. Immodium is an ...	1308853226	4	opiates	2011-06-23 14:20:26	2011	6

Figure 1: Sample DF

To extract meaning from the text in the comments, LIWC was utilized as a text analysis tool. LIWC-22 includes more than 100 pre-built dictionaries that capture people's social and psychological states. Each dictionary includes a list of words, stems, emoticons, and other verbal constructions that have been identified to represent a psychological category of interest. LIWC matches each word in the text to the dictionary words and calculates the percentage of total words in the text that match each of the dictionary categories. [4] After processing the body of the text, LIWC generates 93 features. These features were then added to the original dataset, resulting in a total of 3,845,116 rows and 103 columns.

	User	Body	C	Score	E	Timestamp	Year	Month	I	J	WC	Analytic	Clout	Authentic	Tone
0	[deleted]	at bonnaroo. it smelled like flowers but I did...	1308881233	1	opiates	2011-06-23 22:07:13	2011	2011-06	c21kc5y	t3_g8qw3	17	7.44	1.00	91.95	1.00
1	nvrwastetree	Lol methadone....abiet old school??	1308877576	1	opiates	2011-06-23 21:06:16	2011	2011-06	c21jzhv	t1_c21iap8	5	93.26	50.00	43.37	99.00
2	ttd	Immodium (loperamide) does in fact have activi...	1308861650	1	opiates	2011-06-23 16:40:50	2011	2011-06	c21iap8	t3_i7bwt	138	76.43	23.44	54.89	78.80
3	nvrwastetree	Heard of this as well...	1308854583	1	opiates	2011-06-23 14:43:03	2011	2011-06	c21hdq2	t1_c21h7r4	5	52.71	97.69	1.00	99.00
4	enthreeoh	You're getting placebo effect. Immodium is an ...	1308853226	4	opiates	2011-06-23 14:20:26	2011	2011-06	c21h7r4	t3_i7bwt	112	21.19	5.42	94.05	25.77
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3845111	gamesnstuf	Yes, you're absolutely right. But again, I'd h...	1659067431	1	opiates	2022-07-29 00:03:51	2022	2022-07	ii325di	t1_ii2gf99	139	36.17	7.52	98.30	8.56
3845112	Busy_Background_195	I actually wouldn't wish it on my worst enemy,...	1659067373	1	opiates	2022-07-29 00:02:53	2022	2022-07	ii321hd	t3_w6zr9n	36	1.00	1.38	87.07	1.00
3845113	BoofDontShoot	It doesn't really lower tolerance in my opinio...	1659067340	1	opiates	2022-07-29 00:02:20	2022	2022-07	ii31z8l	t3_wa7tml	46	88.21	4.11	54.89	93.61
3845114	ChampionshipGreat369	Nah it's only 20 just pop it with some weed an...	1659067323	1	opiates	2022-07-29 00:02:03	2022	2022-07	ii31y1f	t3_wan5to	21	2.86	92.33	1.00	95.81
3845115	Psychological_Ad853	He needs to wait until WITHDRAWALS have fully ...	1659067259	1	opiates	2022-07-29 00:00:59	2022	2022-07	ii31to1	t1_ii2z4if	124	34.28	74.07	39.18	1.23

3845116 rows × 103 columns

Figure 2: Combined DF: Comments and LIWC

LIWC includes four summary measures, namely Analytical Thinking, Clout, Authenticity, and Emotional Tone (refer to Fig. 3). These summary measures are derived from standardized scores of various LIWC variables, which are then converted to percentiles based on a normal distribution ranging from 1 to 99.[4] Although these summary features will be utilized in statistical analysis, they will not be used in training the model as they contain information from other LIWC features.

WC	Analytic	Clout	Authentic	Tone
17	7.44	1.00	91.95	1.00
5	93.26	50.00	43.37	99.00
138	76.43	23.44	54.89	78.80
5	52.71	97.69	1.00	99.00
112	21.19	5.42	94.05	25.77

Figure 3: Sample of LIWC values (first 5 features)

To determine whether social media can serve as a predictor of opioid-related deaths, we obtained data from Data.gov to gather information on monthly overdose deaths from 2014 to

2022 (Fig. 4). [1] According to the Centers for Disease Control and Prevention (CDC) Opioids were involved in 68,630 overdose deaths in 2020 (74.8% of all drug overdose deaths). We make an assumption that this is standard rate from 2014-2022. Figure 4 shows the adjusted drug overdose rate with NA values replaced with the average.[3]

	Year	Month	Drug Overdose
0	2014	1	3020.0
1	2015	1	3266.0
2	2016	1	3473.0
3	2017	1	4675.0
4	2018	1	4244.0
...	...	...	...
31	2022	8	6248.0
32	2022	9	4628.0
33	2022	10	4628.0
34	2022	11	4628.0
35	2022	12	4628.0

108 rows × 3 columns

Figure 4

By joining this data with the Opiates subreddit forum dataset, we were able to create an aggregated dataset that included all LIWC values (posts) for each month/year that corresponded to the recorded deaths (Fig. 5).

Body	Score	Timestamp	Year	Month	WC	Analytic	Clout	Drug Overdose
Yuppy Yup!	1	2015-06-23 23:58:16	2015	6	2	93.26	50.00	3132.0
And I would take them 4-5 hours before bed to ...	2	2015-06-23 23:58:00	2015	6	18	10.01	13.32	3132.0
Yeah. I wouldn't worry. Don't make it obvious ...	1	2015-06-23 23:56:43	2015	6	30	8.69	25.24	3132.0
Wow. I'm really happy to hear you're doing wel...	2	2015-06-23 23:56:38	2015	6	295	21.78	44.62	3132.0
I was fine after 5 days.	2	2015-06-23 23:54:42	2015	6	6	62.04	4.80	3132.0

Figure 5

In the next section, we describe the data cleaning and processing steps we took to prepare the data for statistical analysis and model training.

Analysis

Before conducting statistical analysis on the drug overdose data, we performed some preprocessing steps. Since the data is separated by month and year, we aggregated it to obtain a monthly representation of the sentiment using the Linguistic Inquiry and Word Count (LIWC) analysis. As shown in Figure 6, we grouped the data by month and year columns and computed the average LIWC values. This approach allows us to explore the changes in sentiment over time and identify potential patterns or trends.

Although the aggregation process may introduce some level of uncertainty or bias in the data, we made the decision to perform our analysis and models on the average aggregation due to its suitability for our research questions. It is important to acknowledge that summing or averaging the LIWC scores may not accurately capture the variability of sentiment within a given month and may introduce additional noise to the data. Furthermore, the LIWC analysis itself may have some limitations such as the reliance on a predefined dictionary of words and the potential for misclassifying the sentiment of certain phrases or contexts. To address these potential sources of error, we conducted sensitivity analysis by testing different aggregation methods and compared the results. Specifically, we tested both summing and averaging the LIWC values, and found that the average aggregation was a better representation of data.

User	Body	Score	Timestamp	Year	Month	WC	Analytic	Clout	Authentic	Tone	Drug Overdose
i_fucking_love_dope	Yuppy Yup!	1	2015-06-23 23:58:16	2015	6	2	93.26	50.00	1.00	25.77	3132.0
jaynumbernine	And I would take them 4-5 hours before bed to ...	2	2015-06-23 23:58:00	2015	6	18	10.01	13.32	58.07	98.27	3132.0
i_fucking_love_dope	Yeah. I wouldn't worry. Don't make it obvious ...	1	2015-06-23 23:56:43	2015	6	30	8.69	25.24	23.51	25.77	3132.0
momo45678	Wow. I'm really happy to hear you're doing wel...	2	2015-06-23 23:56:38	2015	6	295	21.78	44.62	76.95	80.93	3132.0
BlackQueenCleopatra	I was fine after 5 days.	2	2015-06-23 23:54:42	2015	6	6	62.04	4.80	99.00	99.00	3132.0

Month	Year	Score	WC	Analytic	Clout	Authentic	Tone	Drug Overdose
1	2014	1.825642	34.692438	42.696945	45.342720	48.759547	46.209516	3020.0
2	2014	1.830812	39.261472	42.384175	45.914289	49.320585	45.769275	2921.0
3	2014	1.819408	36.506368	42.969572	44.692573	49.670858	45.749272	3080.0
4	2014	1.861282	39.061421	43.053183	45.250481	49.716061	46.706511	2800.0
5	2014	1.784770	40.287228	43.106703	44.070064	50.527198	45.276802	2944.0

Figure 6

Before model training, some preprocessing steps were undertaken. To perform classification on the target label "Drug Overdose", which is a continuous value, we need to convert the values into discrete classes. In Figure 8, we have a sample set of continuous overdose data represented as monthly values. To bin the data, we can use the following code, which divides the data into classes based on the specified bin ranges. For example, the first class is defined as values between 2000 and 3000. The bin ranges were chosen to predict deaths to the nearest thousand.

Sample Set (continuous): [2800.0, 2811.0, 2837.0, 2921.0, 2926.0, 2944.0, 2945.0, 2990.0, 3002.0, 3016.0, 3020.0, 3065.0, 3080.0, 3132.0, 3149.0, 3184.0, 3211.0, 3266.0, 3346.0, 3361.0, 3369.0, 3373.0, 3410.0, 3437.0, 3473.0, ..., 7004.0, 7061.0, 7081.0, 7100.0, 7118.0]

```
df_avg_class['Drug Overdose'] =  np.digitize(df_avg_class['Drug Overdose'],
                                             bins=[2000,3000,4000,5000,6000,7000,8000])
```

Number of Classes (discrete): [1, 2, 3, 4, 5, 6]

Figure 8: Discretization of class label

In order to perform text clustering of the body data, it is necessary to preprocess the text by removing stop words and splitting the text into individual words. This step is essential to transform the unstructured text data into a structured format that can be used for clustering analysis. Figure 9 displays the code used for preprocessing a sample comment and shows the resulting output after removing the stop words and tokenizing the text.

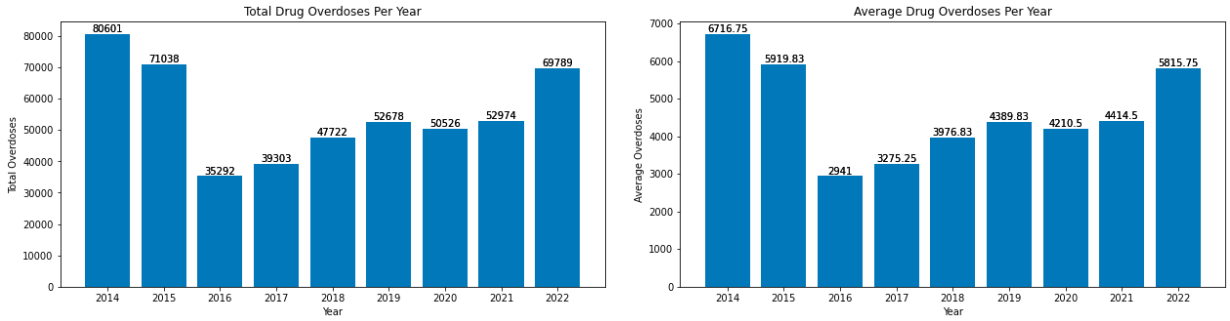
```
import gensim
from gensim.utils import simple_preprocess
import nltk
#nltk.download('stopwords')
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use','https','com','reddit','www','rules','also','amp','sub','know','people','someone','wiki'])
def sent_to_words(sentences):
    for sentence in sentences:
        # deacc=True removes punctuations
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc))
             if word not in stop_words] for doc in texts]
data = df_comments['B']
data_words = list(sent_to_words(data))
# remove stop words
data_words = remove_stopwords(data_words)
```

```
'Just get some food in your stomach and drink as much water as you can.  Sleep it off.'

['get', 'food', 'stomach', 'drink', 'much', 'water', 'sleep']
```

Figure 9

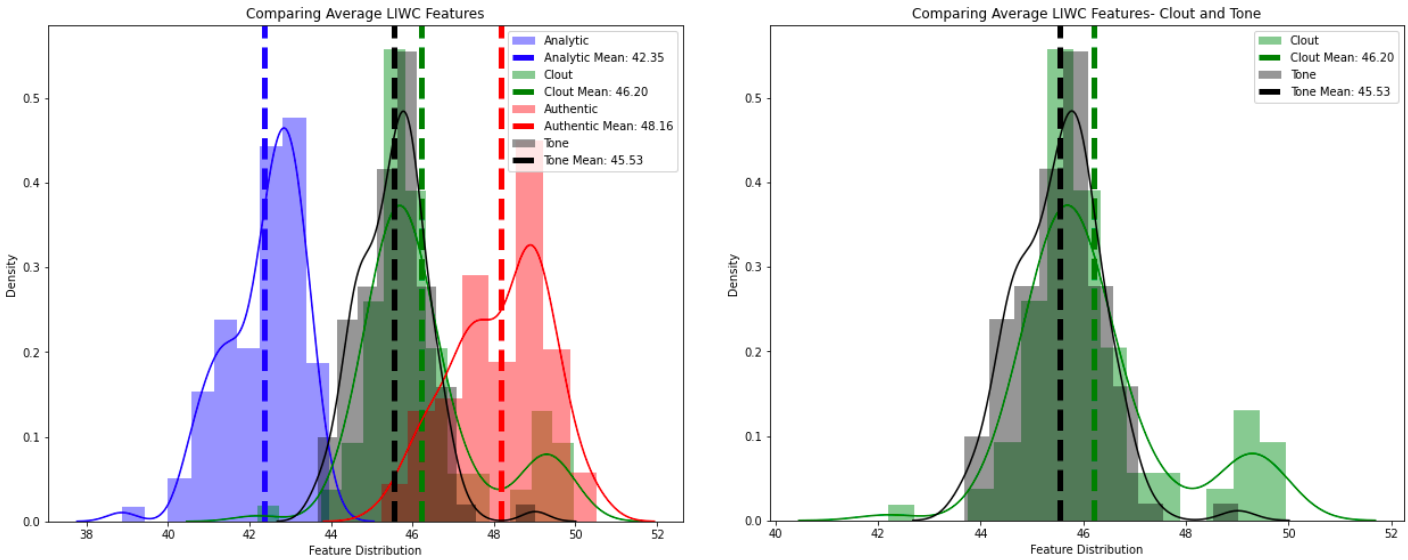
After processing the data, we are now ready to conduct our analysis and train our models. The response variable we are evaluating is the number of drug overdose incidents in a given month. In order to gain a better understanding of the data, we examined the average and total deaths per year. As shown in Figure 10, there was a significant drop in overdose incidents in 2016. While the number of deaths has remained relatively consistent from 2018-2022, it's worth exploring potential reasons for the drop in 2016. One possibility is the increased public awareness and education campaigns surrounding the opioid epidemic. Another possibility is increased access to naloxone, a medication used to counteract the effects of opioids, which can reduce the likelihood of overdose deaths.[14] In addition to investigating these potential factors, we also examined the distributions of the LIWC average values and our selected features.



(a) (b)

Figure 10: Overdoses in U.S. 2014-2022

As described above LIWC has 4 average variables: Analytical Thinking, Clout, Authenticity, and Tone. Analytical Thinking measures the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns. People low in Analytical Thinking tend to use language that is more intuitive and personal, while those high in Analytical Thinking tend to be rewarded in academic settings and are correlated with things like grades and reasoning skills. Clout refers to the relative social status, confidence, or leadership that people display through their writing or talking. Authenticity measures the degree to which people are self-monitoring when they reveal themselves in an "authentic" or honest way. Finally, the Tone variable puts the Positive and Negative Tone dimensions into a single summary variable, with numbers above 50 indicating a more positive tone and numbers below 50 suggesting a more negative emotional tone. [4] The comparison of the LIWC values in Figure 11 sheds light on the overall sentiment of the text in the analyzed forums. The similar mean values of Analytic, Clout, Authentic, and Tone suggest that the posts are evenly distributed across these four writing styles. However, the isolated plots of Clout and Tone in Figure 11b demonstrate that both are fairly normally distributed. The highest found attribute in the text is Authenticity, while Analytical Thinking is the lowest. This aligns with the nature of the data being self-help or advice-related, which may not require much logical or analytical thinking. Furthermore, the distribution tone indicates that the comments are evenly balanced in positive and negative sentiment. Overall, this analysis provides valuable insights into the sentiment of the analyzed forum posts, highlighting the importance of honesty and seeking help as major themes.



(a) (b)

Figure 11: Average LIWC value analysis

As part of our model training preparation, we carefully selected the features shown in Figure 12 that capture the majority of sentiment in the body text. We excluded syntax-related features such as word count and number of commas, as our goal is to model the sentiment of the comments and understand how it relates to real-world sentiment.

pronoun	drives	sad	home
ppron	affiliation	social	money
adverb	achieve	family	relig
negate	power	friend	death
verb	reward	health	informal
adj	risk	sexual	swear
anx	work	ingest	netspeak
anger	leisure	drives	assent

Figure 12: Feature selection

To assess feature correlations, we first used VIF analysis (Figure 13), which indicates the degree of correlation between an explanatory variable and others in the model. Values greater than 5 suggest potentially severe correlation. As most VIF values were above 5, we performed further analysis using factor analysis to find groupings of features and form factors. [12]

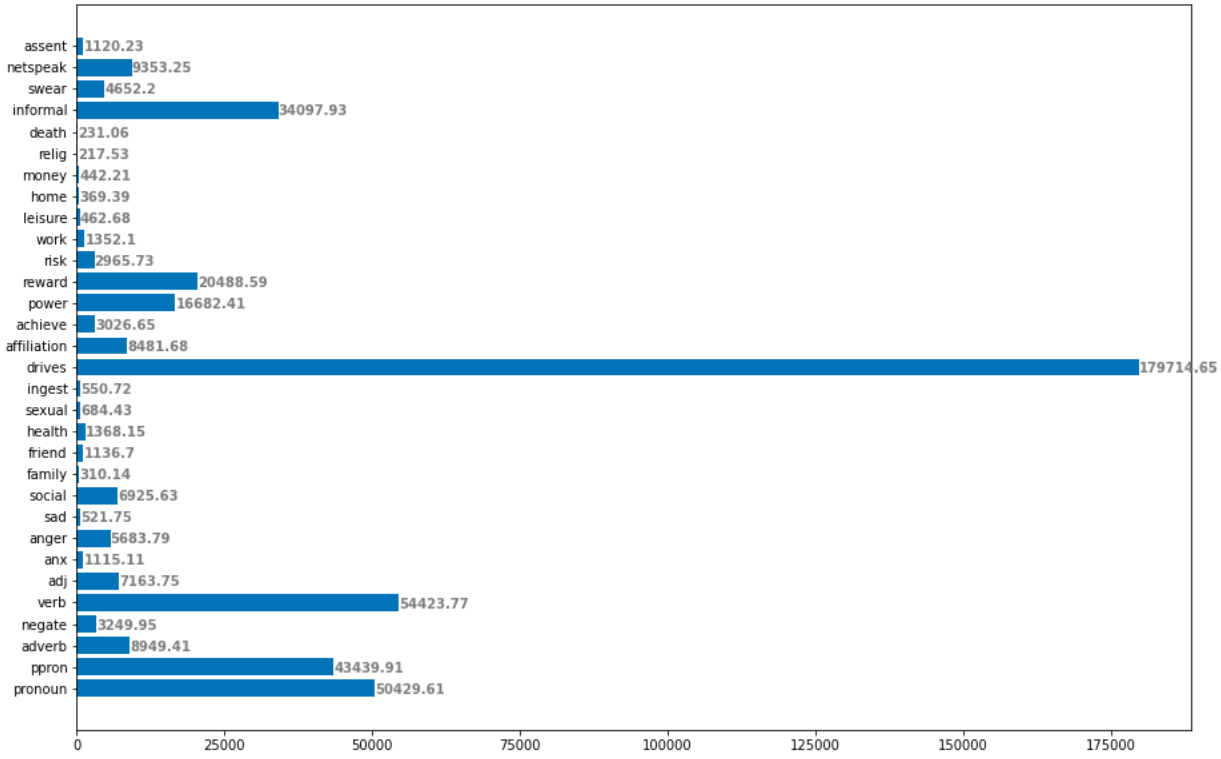


Figure 13

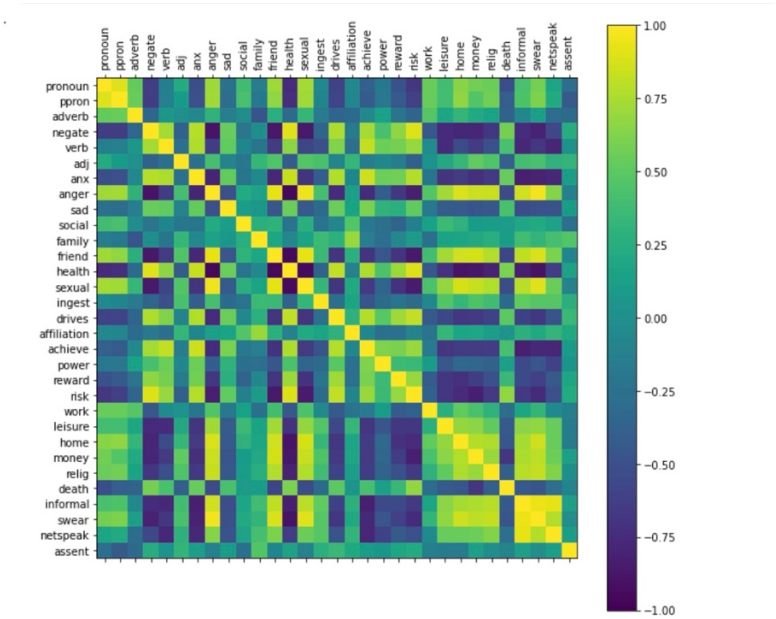


Figure 14: Correlation Matrix (For details refer to Appendix B)

Bartlett's test confirmed the presence of correlation, while the KMO measure indicated good suitability for factor analysis. Using the Scree Plot (Fig. 15), we fit 5 factors for our data, with loadings indicating the degree of factor influence on variables. [2]

Based on our analysis from Table 1, we can identify some potential factors for these features, and then decide which of these factors should be used for training the model. The first

group of features, Netspeak, Swear, and Leisure, may potentially form a factor related to informal language use. The second group of features, Pronoun and Ppron, may represent the degree of self-reference in the text. The third group of features, Affiliation, Social, and Family, may be related to the social context in which the text was written. Finally, the last group of features, Verb, Risk, Death, and Reward, may represent the topic or content of the text. We can use the correlation matrix in Figure 14 to confirm these findings.

```
Chi squared value :  5208.505603129671
p value :  0.0
KMO Model: 0.8585611912104225
```

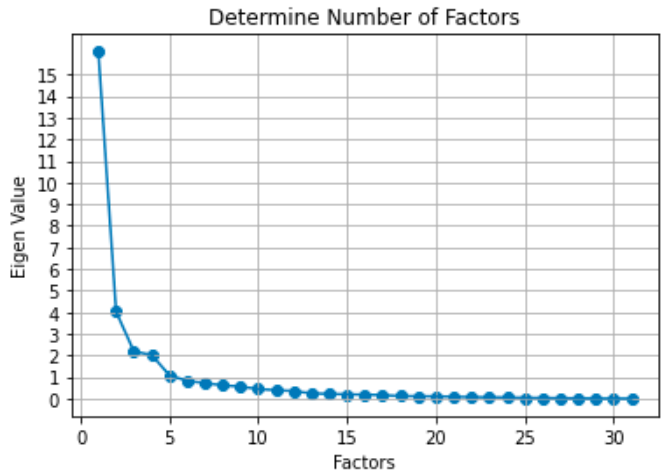


Figure 15

	0	1	2	3	4
netspeak	0.884776	0.040117	0.224192	0.194173	0.195038
informal	0.839116	0.345167	0.307697	0.169830	0.170645
swear	0.720238	0.567439	0.268213	0.152205	0.139603
leisure	0.632404	0.368118	0.010735	0.280154	-0.370591
anger	0.631196	0.723997	0.182655	0.086129	0.013798
friend	0.591599	0.701735	0.202715	0.208844	-0.113384
relig	0.566812	0.555200	0.288600	0.046908	0.190064
money	0.548874	0.595466	0.400613	0.042721	-0.236389
sexual	0.522097	0.762783	0.208013	0.167242	-0.073346
home	0.517832	0.705026	0.125194	0.161705	-0.044366
ingest	0.502702	-0.041335	0.502296	-0.021161	-0.123892
ppron	0.208908	0.884462	-0.261989	0.049000	0.201454
family	0.197978	-0.182514	0.527498	0.642832	0.231345
pronoun	0.161102	0.898764	-0.110814	0.048417	0.054047
adj	0.131960	0.284947	0.699677	0.069435	-0.091709
affiliation	0.104048	-0.070657	0.310917	0.857404	-0.169828
social	0.089582	0.272468	-0.417013	0.744818	0.066641
work	-0.001378	0.643946	-0.095393	-0.121324	-0.091583
assent	-0.065428	-0.248494	0.628499	0.063012	0.081453
adverb	-0.156696	0.627140	-0.192503	-0.285087	-0.223932
death	-0.308183	-0.488879	-0.280446	0.212686	0.312075
reward	-0.503748	-0.483814	0.086754	-0.241963	0.300677
sad	-0.574657	-0.223006	-0.044195	0.269338	0.055281
risk	-0.592928	-0.640220	-0.070317	0.042201	0.346521
negate	-0.635877	-0.631081	-0.061149	-0.056113	0.276324
health	-0.652707	-0.722908	-0.135178	-0.043880	0.088212
power	-0.659205	0.001478	0.063321	-0.228827	-0.026315
drives	-0.701770	-0.528917	0.219718	0.278340	0.149249
anx	-0.822082	-0.378117	-0.032937	-0.013019	0.065473
verb	-0.876215	-0.052698	-0.124333	-0.102634	0.368360
achieve	-0.907415	-0.244838	0.024686	-0.006023	0.052291

Table 1: Factors

The goal of this analysis is to determine which factors or groupings of variables should be included in the model training process to improve the understanding of sentiment in the comments. The first group of features, Netspeak, Swear, and Leisure, may potentially form a factor related to informal language use. The second group of features, Pronoun and Ppron, may represent the degree of self-reference in the text. The third group of features, Affiliation, Social, and Family, may be related to the social context in which the text was written. Finally, the last group of features, Verb, Risk, Death, and Reward, may represent the topic or content of the text. Based on the results of our analysis, we have identified potential factors related to informal language use, self-reference, and social context, which we believe are relevant for this purpose. Therefore, we will consider including these factors in the model and evaluate their performance. We will also test the model's performance by removing certain features and factors to determine which combination yields the best results.

By analyzing the distributions of the selected features in Figure 16, we can identify potential areas for further exploration and tuning of our model. It is important to note that the distributions of features can greatly impact the performance of machine learning models. When a feature has a larger variance, it may be more influential in predicting the target variable, and thus may require more attention during model training. On the other hand, features that are closer to normal may require less tuning and may not have as significant of an impact on the model's performance.

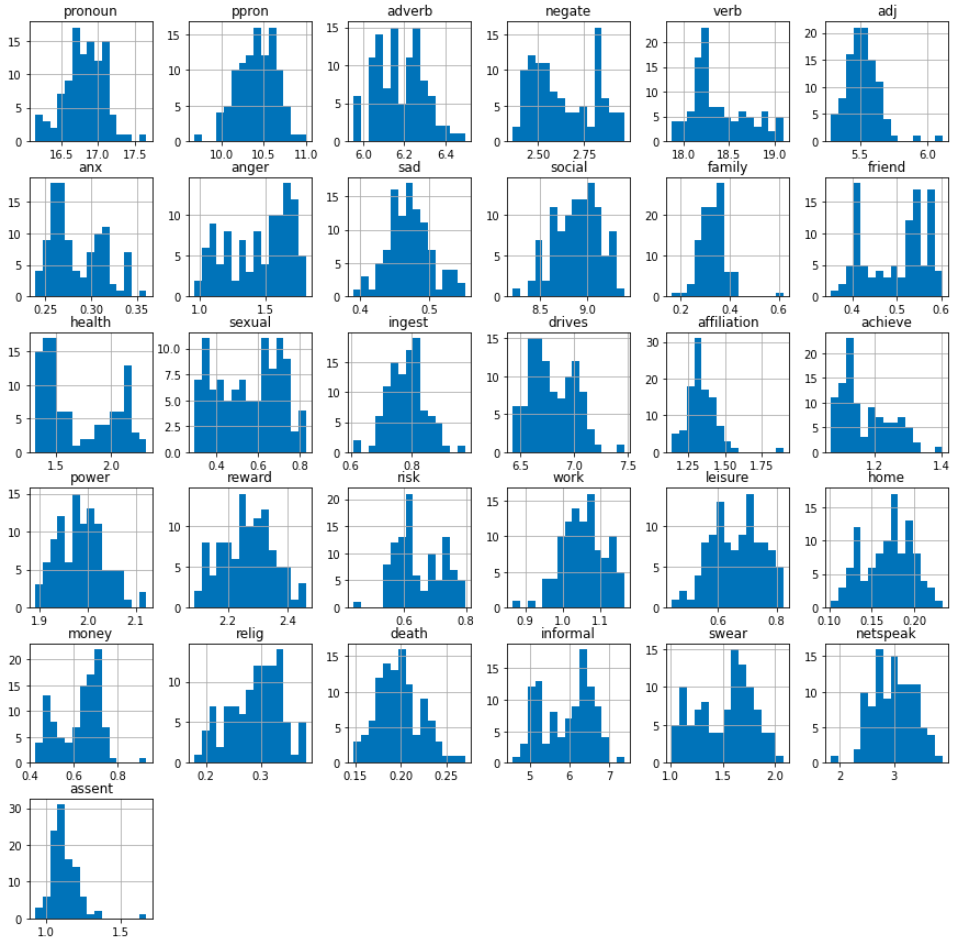


Figure 16: Distributions of Selected Features

The next section of the paper aims to evaluate the performance of the Classification and Regression models trained on the dataset, as well as explore the insights obtained from topic clustering. The goal is to determine how well the models perform in predicting drug overdoses, and to gain a deeper understanding of the different topics that emerged from the Reddit comments.

Model Development and Application of model(s)

Table 2 displays the features used for training the Classification and Regression models. The models were trained with the target label being Drug Overdoses. For the Classification model, the goal was to predict overdoses to the nearest thousand due to the low number of rows in our dataset. This allowed for simpler prediction and still provided insights into the overall

deaths for a given month. In contrast, Multivariate Regression was attempted to see if the data was large enough to predict the continuous value of overdoses.

To compare the results, 3-4 different models were trained for each technique. The following section will discuss how well the models performed. In addition to the Classification and Regression models, topic clustering was also done to identify the different topics in the Reddit comments. By examining the performance of the models and the insights obtained from topic clustering, we hope to gain a better understanding of the factors that contribute to drug overdoses and inform future research in this area.

pronoun	ingest
ppron	drives
adverb	affiliation
negate	achieve
verb	power
adj	reward
anx	risk
anger	work
sad	leisure
social	home
family	money
friend	relig
health	death
sexual	netspeak
ingest	assent

Table 2: Features

Classification

We chose three different models for performing classification: Random Forest (RF)[1], Gradient Boosting Classifier (GB)[2], and Extreme Gradient Boosting Classifier (XGB)[3].

These classification models are known for their ability to produce accurate results. One important factor when tuning these models is selecting the number of estimators for the classifiers to run. Figure 17 displays a plot that shows how each model performs at different estimator values. Based on the plot, we will choose an estimator value of 100 because both XGB and GB classifiers perform equally well after 100, while the RF model performs worse.

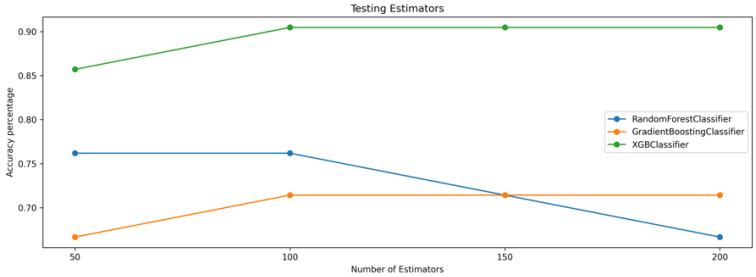


Figure 17: N_estimators

[1]Random Forest is an ensemble method that creates multiple decision tree classifiers on different subsets of the dataset and combines them by averaging the results to improve accuracy and reduce overfitting. [9]

[2]Gradient Boosting Classifier is an algorithm that builds an additive model in a step-by-step manner and allows for the optimization of differentiable loss functions to improve the accuracy of the model. [8]

[3]XGBoost is a highly optimized, distributed gradient boosting library that is designed to be efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework and uses parallel processing to create decision trees, which results in faster and more accurate predictions. [11]

During the model training phase, the dataset was split into an 80-20 train-test split, and each of the selected models was run on the dataset to measure its prediction accuracy. As shown in Table 3, the XGB model outperformed the other models, with an accuracy of approximately 90 percent. However, due to the small size of the dataset, the high accuracy results may be misleading, and the model is likely to overfit. To mitigate this issue, we performed cross-validation to evaluate the models' performance. Given the small dataset size, we chose Leave-One-Out Cross-Validation (LOOCV) over K-Fold, a technique where we test the model with a single sample left out while training on the remaining samples.

Comparing the LOOCV accuracy scores in Table 3 to the measured accuracy scores, we can observe that LOOCV performs much closer to the actual accuracy. The RF and GB models perform well using LOOCV, with the RF model achieving approximately 76 percent accuracy and the GB model approximately 71 percent accuracy. Although the XGB model is the best performing algorithm, as seen in Table X, it can lead to overfitting, as shown in the cross-validation row. This can be attributed to the complex structure of the XGB algorithm and its tendency to overfit the training data if the hyperparameters are not tuned properly or if the model complexity is too high.

Evaluation Metric	RandomForestClassifier	GradientBoostingClassifier	XGBClassifier
Accuracy Score	0.7619047619047619	0.7142857142857143	0.9047619047619048
Cross Validation Average Score-LOOCV	0.7788461538461539	0.7115384615384616	0.6923076923076923
Cross Validation Average Score-5 K Fold	0.6914285714285715	0.6142857142857143	0.5961904761904762

Table 3: Classification Model Evaluation

One of the benefits of performing classification and Random Forest is the ability to see the importance of features in making predictions. By examining feature importance, we can identify which features are the most influential in fitting the model. The Table 4 displays the importance of features ranked from the most important to the least. For instance, features such as family, netspeak, and negate have the highest importance.

The family feature reflects the frequency of family-related words in the text, indicating a stronger support system and potentially reducing the risk of drug overdose. Netspeak measures the frequency of online communication jargon, which can suggest increased exposure to drug culture and higher risk of drug overdose. Negate reflects the frequency of negative sentiment in the text, potentially indicating greater risk for drug overdose due to mental health issues and substance use disorders.

The SHAP (SHapley Additive exPlanations) allow us to understand the contribution of each input feature in predicting a certain class, compared to a baseline prediction. The plot in Figure 18 illustrates the impact of each feature on the predicted class. The length of the bars for each feature indicates the strength of the impact, while the color shows the class of the impact. For instance, a higher frequency of family-related words in the text, as indicated by the family feature, increases the probability of predicting class 2. The SHAP values provide insights into which features are most significant in predicting the target variable and help to interpret the model's output. [6]

	Features	Importance
0	family	0.076429
1	netspeak	0.058099
2	negate	0.053022
3	adverb	0.052474
4	health	0.048233
5	sexual	0.047107
6	risk	0.045304
7	anger	0.043810
8	swear	0.042698
9	drives	0.042145
10	work	0.037732
11	informal	0.037380
12	money	0.036827
13	home	0.036624
14	pronoun	0.030470
15	death	0.029421
16	anx	0.029332
17	affiliation	0.027517
18	relig	0.026071
19	leisure	0.023517
20	ppron	0.023084
21	assent	0.020507
22	friend	0.019196
23	achieve	0.017891
24	ingest	0.016725
25	sad	0.014922
26	power	0.014443
27	reward	0.013144
28	adj	0.012610
29	social	0.011656
30	verb	0.011607

Table 4: RF Feature Importance

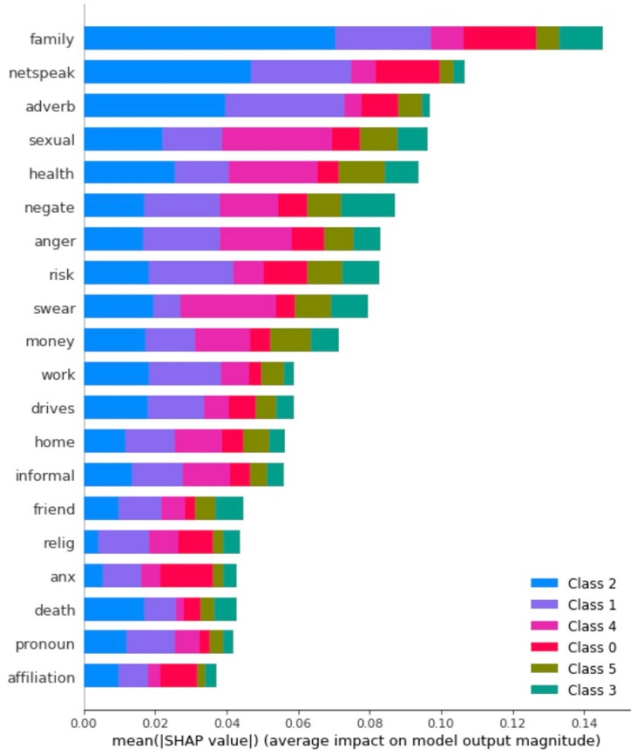


Figure 18: SHAP Tree Explainer

Now that we have explored the results of our classification models, we will analyse regression models and see how well they performed in predicting continuous values of overdoses.

Multivariate Regression

Based on the promising results from the classification analysis, we decided to investigate whether regression analysis could be applied to predict the continuous values of drug overdoses for a given month. The data preprocessing steps were similar to the classification analysis, and we used LOOCV for evaluation due to the size of our dataset and the tests conducted above.

To perform the regression analysis, we trained four different models out of multiple that were tested. The first model was a simple Multivariate Linear Regression to see how a basic algorithm works on our dataset. Next, we trained Ridge and Lasso models, which are both regularized linear regression techniques. Ridge regression is a method that adds a penalty term to the cost function to avoid overfitting, whereas Lasso regression performs variable selection by shrinking some of the coefficients to zero. These two algorithms are beneficial for our dataset because they can handle multicollinearity and prevent overfitting which is beneficially given our features and the small nature of the dataset. Finally, we used XGBRegressor, which was discussed above.

Two metrics were used to evaluate the regression models - R-squared score and Mean Squared Error (MSE). While R-squared measures the proportion of total variation in the dependent variable explained by the independent variables, MSE measures the average squared difference between predicted and actual values, indicating the model's prediction error. (Refer to

images in Appendix C) [13] Looking at Table 5, we can see that Linear and XGB have the highest R-squared scores, indicating a good fit to the data. However, the issue is that the MSE values and cross-validation scores suggest that these models may be overfitting, as some of the predicted values are close to the trendline but with large MSE. The CV MSE is much greater than the calculated MSE from model testing, which further supports the presence of overfitting. In contrast, Ridge and Lasso also have good R-squared scores, but with less overfitting due to the regularization nature of these models.

Metrics	LinearRegression()	RidgeCV(alphas=array([0.1, 1. , 10.]), cv=5)	LassoCV(cv=5)	XGBRegressor()
R^2 Score	0.9154843531345928	0.8948055999265808	0.907290343200958	0.9163068073231104
MSE	147224.62555340317	183246.61450899945	161498.43270039663	145791.92623167663
Cross Val MSE Avg	330155.2250258535	285421.7931919859	323124.52273181203	264124.45000954776

Table 5: Evaluation of Regression Model

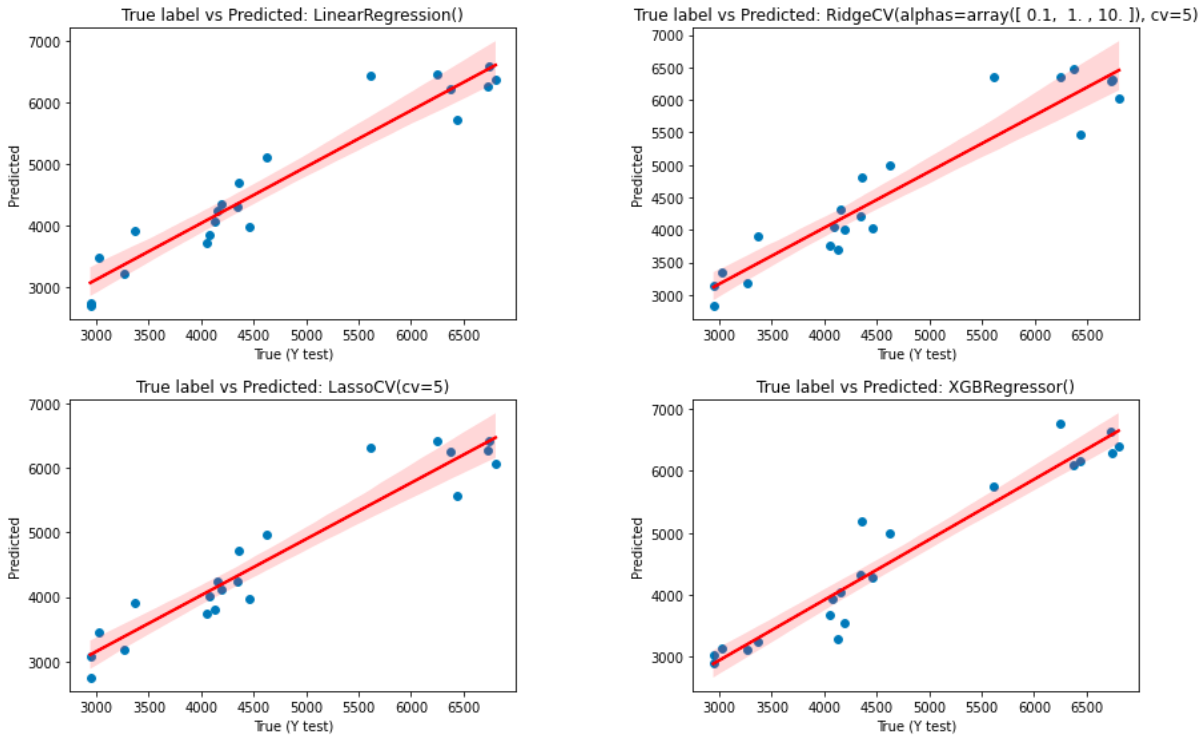


Figure 19: Plots of True vs Predicted

Based on the observations above, to mitigate overfitting, it would be beneficial to explore additional methods such as reducing the number of features or hyperparameter tuning. Despite the limitations posed by the number of data points, the regression analysis yielded decent accuracy, but more data and further model tuning may be required to build a more robust regression model. Therefore, any predictions made using the current model should be taken with caution and not as absolute fact.

Now that we have evaluated the performance of our regression models, let's move on to the next step of our analysis, which is topic clustering.

Topic Clustering

In this paper, one of our main objectives is to identify the sentiment of the text by analyzing different topics in the comments section. To achieve this, we conducted text cleaning on the Body column and applied Latent Dirichlet Allocation (LDA) for topic clustering. We then used the LDAvis visualization tool to represent the topics as circles, with the size of the circles indicating their prevalence, and the inter-topic distances revealing their similarity or dissimilarity. Additionally, LDAvis provides a panel of words associated with each topic, with their size indicating their relevance to that particular topic. This interactive visualization allows users to explore and interpret the topics and associated words. The resulting visualization is displayed in Figure 20, with an interactive version available on our Github repository. [7]

Looking at Figure 20 and the associated words, we can see that the topics are related to drug usage, pain management, and general conversations. For example, topic 4 has words related to drug usage such as "get", "got", "dope", "money" etc. Similarly, topic 3 has words related to the dosage of drugs such as "mg", "oxy", "dose", "tolerance" etc. Topic 1 seems to be more about pain management with words like "pain", "feel", "days" etc. Based on the identified words in each topic, it is possible to group them into larger categories. For instance, topics 3 and 4 can be grouped as drug addiction, whereas topic 1 can be grouped as pain management. Overall, the LDA topic model provides a way to understand the underlying themes and topics present in the comments, which can aid in gaining insights and improving the understanding of the data.

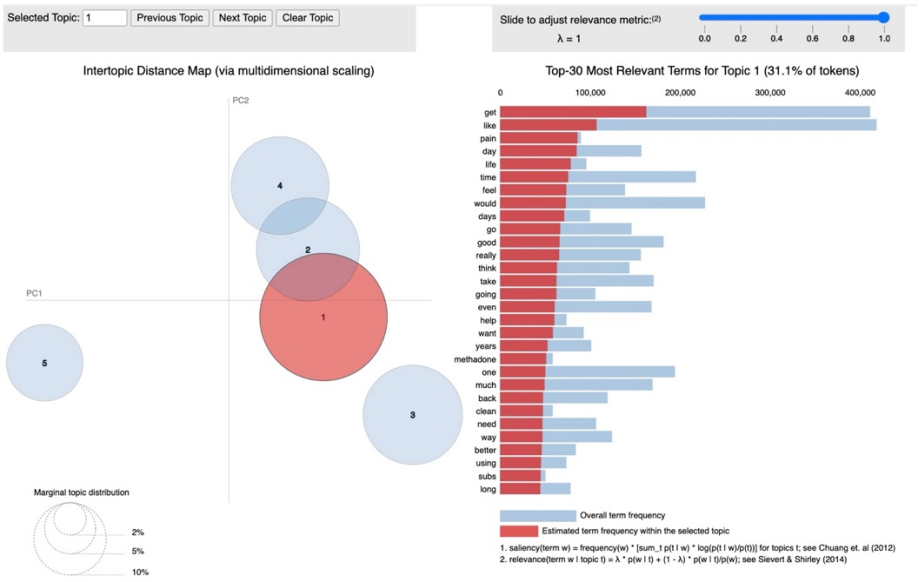


Figure 20: Sample of Topic Clustering Output

Conclusions and Discussion

In this paper, we started by cleaning our dataset and extracting the LIWC from the body text. We then performed statistical analysis, including Factor Analysis and examining the distributions of different variables, which allowed us to make initial assumptions and select models. We utilized three machine learning techniques, including classification, multivariate regression, and topic clustering, and observed positive results for classification and topic clustering, while regression needs further tuning and data collection. Throughout the project, we

have changed our analysis and modeling techniques as we found new insights into the data. Originally, we used PCA and clustering analysis but found that factor analysis was a better fit for our data. We tried different models and evaluation techniques, and in hindsight, a better process for choosing and tuning our models would have been beneficial.

Our analysis can model the sentiment of a given forum, and using this and the idea found in research, we can suggest that these Reddit forums can be used as a sensor for opioid-related deaths. While we cannot state our results as fact, we believe they have enough merit to provide a rough estimate of deaths in a given month to the nearest thousand. Further work needs to be done in feature engineering, model tuning, and data collection, specifically in examining the correlations between different features and selecting the best features for better results. We also need to tune our models more to reduce over/under-fitting and increase accuracy. Finally, we note that our data only spans from 2014 to 2022 and accessing a larger range of data could lead to more successful results.

References

[1] "Data Catalog." *Catalog*, https://catalog.data.gov/dataset/?tags=drugoverdose&res_format=CSV.

[2] Babu, Dhamodaran. "Dimensionality Reduction Using Factor Analysis in Python!" *Analytics Vidhya*, 29 Dec. 2020, https://www.analyticsvidhya.com/blog/2020/10/dimensionality-reduction-using-factor-analysis-in-python/.

[3] "Death Rate Maps & Graphs." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 2 June 2022, https://www.cdc.gov/drugoverdose/deaths/index.html.

[4] *LIWC*, https://www.liwc.app/help/liwc.

[5] "Opiates (U/Opiates)." *Reddit*, https://www.reddit.com/user/opiates.

[6] "Shap.treeexplainer¶." *Shap.TreeExplainer - SHAP Latest Documentation*, https://shap-lrjball.readthedocs.io/en/latest/generated/shap.TreeExplainer.html.

[7] Sievert, Carson, and Kenneth Shirley. "LDAvis: A Method for Visualizing and Interpreting Topics." *ACL Anthology*, https://aclanthology.org/W14-3110/.

[8] "Sklearn.ensemble.gradientboostingclassifier." *Scikit*, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html.

[9] "Sklearn.ensemble.randomforestclassifier." *Scikit*, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[10] Team, DataCamp. "Lasso and Ridge Regression in Python Tutorial." *DataCamp*, DataCamp, 25 Mar. 2022, https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression.

[11] "XGBoost Documentationℑ." *XGBoost Documentation - Xgboost 1.7.5 Documentation*, https://xgboost.readthedocs.io/en/stable/.

[12] Zach. "How to Calculate VIF in Python." *Statology*, 12 Oct. 2022, https://www.statology.org/how-to-calculate-vif-in-python/.

[13] "Mean Squared Error or r-Squared - Which One to Use?" *Data Analytics*, 4 Apr. 2023, https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/#Differences_Mean_Square_Error_vs_R-Squared.

[14] Commissioner, Office of the. "FDA Approves First over-the-Counter Naloxone Nasal Spray." *U.S. Food and Drug Administration*, FDA, https://www.fda.gov/news-events/press-announcements/fda-approves-first-over-counter-naloxone-nasal-

spray#:~:text=Narcan%20nasal%20spray%20was%20first,2015%20as%20a%20prescription%20drug.

Literature Review Citations:

Young, S. D. (2014). Behavioral insights on big data: using social media for predicting biomedical outcomes. *Trends in microbiology*, *22*(11), 601-602.

Lo-Ciganic, W. H., Donohue, J. M., Yang, Q., Huang, J. L., Chang, C. Y., Weiss, J. C., ... & Gellad, W. F. (2022). Developing and validating a machine-learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states: a prognostic modelling study. *The Lancet Digital Health*, *4*(6), e455-e465.

Gaikar, D. D., Marakarkandy, B., & Dasgupta, C. (2015). Using Twitter data to predict the performance of Bollywood movies. *Industrial Management & Data Systems*.

Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2018). Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, *44*(3), 408-421.

Guidi, M. (2017). Extracting information from social media to track financial markets.

Barakos, M. (2015). *Social Media and Forecasting: What is the potential of Social Media as a forecasting tool?* (Bachelor's thesis, University of Twente).

Curtis, B., Giorgi, S., Buffone, A. E., Ungar, L. H., Ashford, R. D., Hemmons, J., ... & Schwartz, H. A. (2018). Can Twitter be used to predict county excessive alcohol consumption rates?. *PloS one*, *13*(4), e0194290.

Friedman, J. R., & Hansen, H. (2022). Evaluation of increases in drug overdose mortality rates in the US by race and ethnicity before and during the COVID-19 pandemic. *JAMA psychiatry*, *79*(4), 379-381.

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, *117*(19), 10165-10171.

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (Vol. 1, pp. 492-499). IEEE.

Rousidis, D., Koukaras, P., & Tjortjis, C. (2020). Social media prediction: a literature review. *Multimedia Tools and Applications*, *79*(9-10), 6279-6311.

Matero, M., Giorgi, S., Curtis, B., Ungar, L. H., & Schwartz, H. A. (2023). Opioid death projections with AI-based forecasts using social media language. *NPJ Digital Medicine*, *6*(1), 35.

Sarker, A., Gonzalez-Hernandez, G., Ruan, Y., & Perrone, J. (2019). Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA network open*, *2*(11), e1914672-e1914672.

Appendix A: Literature Review

In this brief commentary, Young (2014) discusses the potential of social media data for predicting biomedical outcomes using behavioral insights. The author highlights the importance of incorporating behavioral data into predictive models, and suggests that social media data could be a valuable source of this information. The article provides a general overview of the topic and offers suggestions for future research.

Lo-Ciganic et al. (2022) describe the development and validation of a machine learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states. The study used administrative claims data and machine learning techniques to identify patients at high risk of overdose, with the goal of providing targeted interventions to reduce the risk of overdose. The study's findings suggest that machine learning algorithms can be used to accurately predict opioid overdose and could be a useful tool for improving patient outcomes.

Gaikar et al. (2015) explore the use of Twitter data to predict the box office performance of Bollywood movies. The authors collected and analyzed tweets about Bollywood movies and used machine learning techniques to predict box office revenue. The study's findings suggest that Twitter data can be used to accurately predict the box office performance of Bollywood movies, and could be a useful tool for the movie industry.

Elshendy et al. (2018) investigate the use of four different online media sources to forecast crude oil prices. The authors collected and analyzed data from social media, news articles, blogs, and forums, and used machine learning techniques to predict crude oil prices. The study's findings suggest that online media data can be used to accurately predict crude oil prices and could be a valuable tool for financial analysts.

In this review article, Guidi (2017) discusses the use of social media data to track financial markets. The author provides an overview of the research in this area, highlighting the potential of social media data for predicting stock prices, market sentiment, and other financial indicators. The article provides a useful summary of the current state of the field and identifies potential areas for future research.

Barakos (2015) presents a bachelor's thesis exploring the potential of social media as a forecasting tool. The author reviews the literature on social media forecasting and conducts a case study to examine the use of Twitter data to predict box office revenue for a movie. The study's findings suggest that social media data can be a useful tool for forecasting

This article by Friedman and Hansen (2022) examines drug overdose mortality rates in the United States by race and ethnicity before and during the COVID-19 pandemic. The authors use data from the Centers for Disease Control and Prevention to analyze trends in overdose deaths across demographic groups. The study found that drug overdose mortality rates increased significantly for all racial and ethnic groups during the pandemic, with the largest increase among Black individuals. The authors suggest that targeted interventions are needed to address these disparities.

Jaidka et al. (2020) investigate the use of Twitter data to estimate subjective well-being across different geographic regions using dictionary-based and data-driven language methods. The study compared different approaches to analyzing social media data and found that data-driven language methods were more effective than dictionary-based methods at estimating subjective well-being in different geographic areas. The authors suggest that their findings have implications for understanding how social media can be used to measure well-being at a local level.

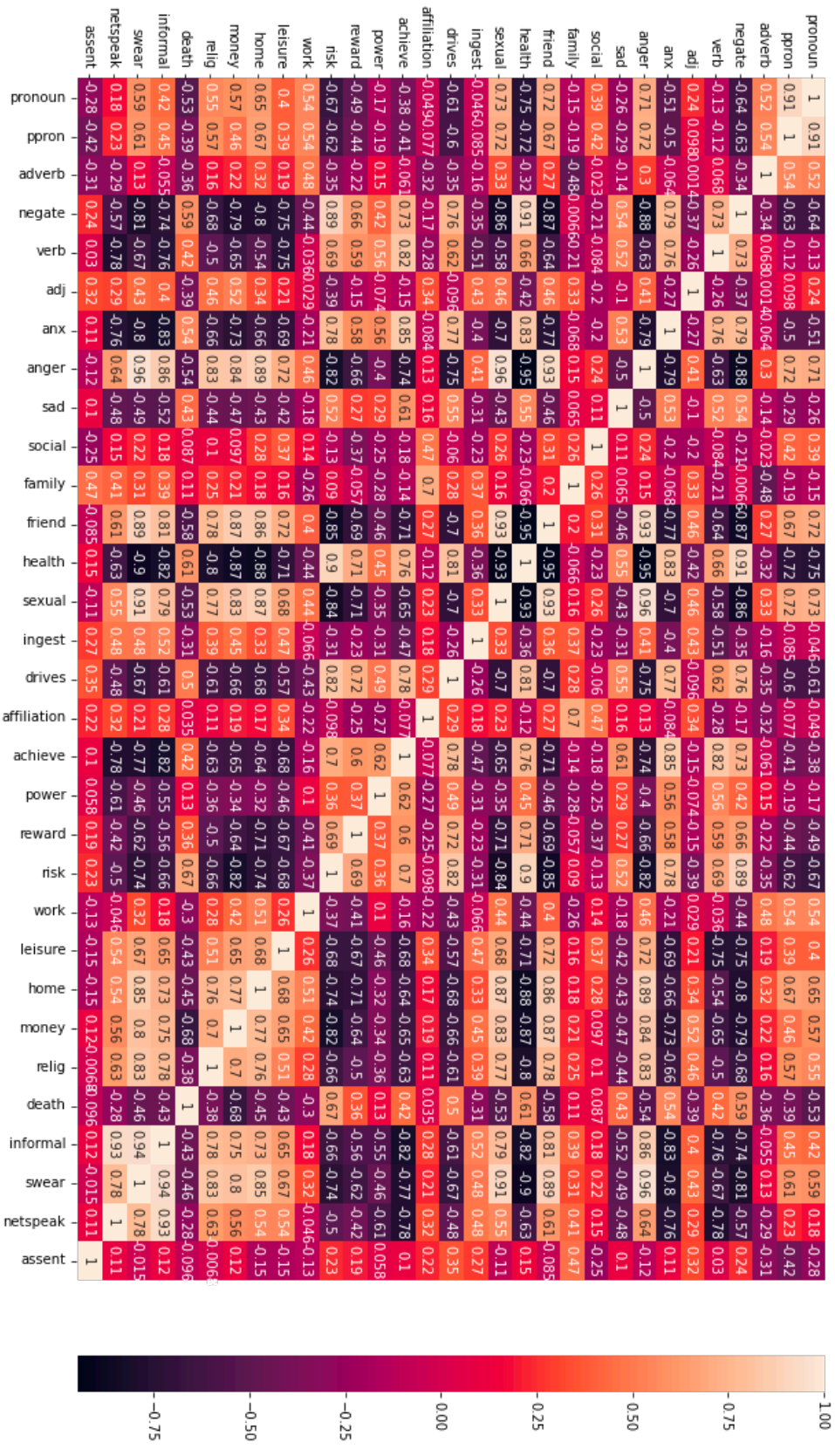
Asur and Huberman (2010) discuss the potential for social media data to be used for predicting future events. The authors describe a study in which they used Twitter data to predict the box office success of movies with high accuracy. They suggest that social media data can be a valuable source of information for predicting trends and events, and that this approach could be applied to a wide range of fields.

Rousidis et al. (2020) present a literature review of studies that use social media data for prediction tasks. The authors review studies across various fields, including politics, health, and economics, and highlight the potential benefits and challenges of using social media data for prediction tasks. They suggest that social media data can be a valuable source of information for predicting events and trends, but that careful consideration of ethical and privacy concerns is necessary.

Matero et al. (2023) use AI-based forecasts and social media language to project opioid overdose deaths in the United States. The authors trained a deep learning model on social media data and used it to make forecasts for opioid overdose deaths. They found that their approach was more accurate than traditional statistical models, and suggest that their findings could have implications for developing targeted interventions to prevent opioid overdose deaths.

Sarker et al. (2019) present a study on the use of machine learning and natural language processing techniques to monitor and characterize opioid-related social media chatter in specific geographic locations. The authors collected Twitter data from users in Philadelphia, Pennsylvania, and applied machine learning algorithms to classify opioid-related tweets and identify trends in the data. The study found that this approach was effective in identifying key topics and concerns related to opioids in the targeted location. The authors suggest that their approach could be used to track opioid-related social media discussions in other locations, providing valuable insights for public health efforts. The study demonstrates the potential of social media data analysis for monitoring public health concerns and identifying emerging trends.

Appendix B: Detailed Correlation Matrix



Appendix C

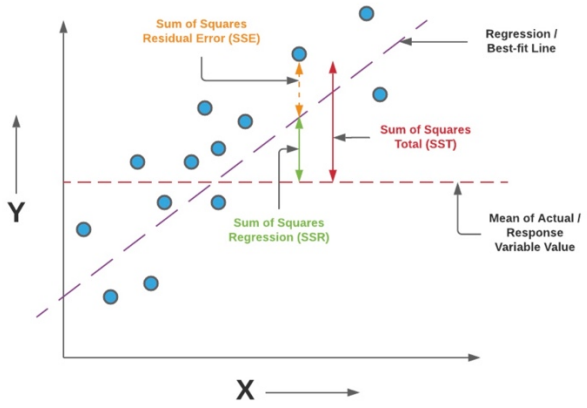


Fig 4. Diagrammatic representation for understanding R-Squared

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

R squared [13]

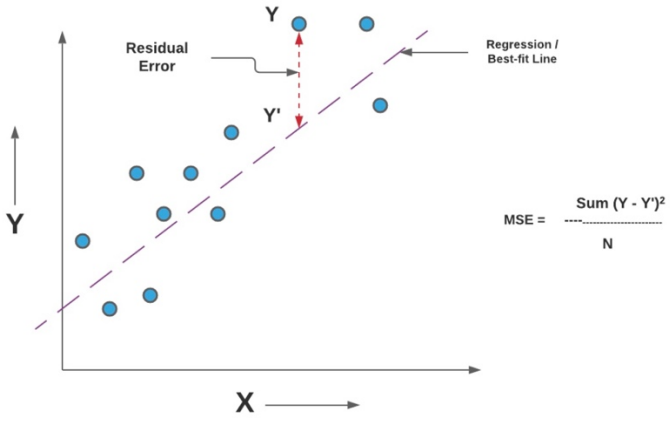


Fig 2. Mean Squared Error Representation

Mean Squared Error [13]