The picture can't be displayed.

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

Noah Kader / kadern@pri.edu
https://github.com/nkader12/DAProject2023
Rensselaer Polytechnic Institute, Tetherless World Constellation, Troy, NY, United States,

Abstract

The opioid crisis has led to a global public health emergency with opioid overdose causing the death of thousands of individuals annually. To address this issue, 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, regression, principal component analysis (PCA), and clustering, to gain insights from the data.

To gain sentiment from the text, we leveraged Linguistic Inquiry and Word Count analysis (LIWC) on the comments' text 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 in a particular area. Additionally, topic clustering proved useful in grouping the 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.

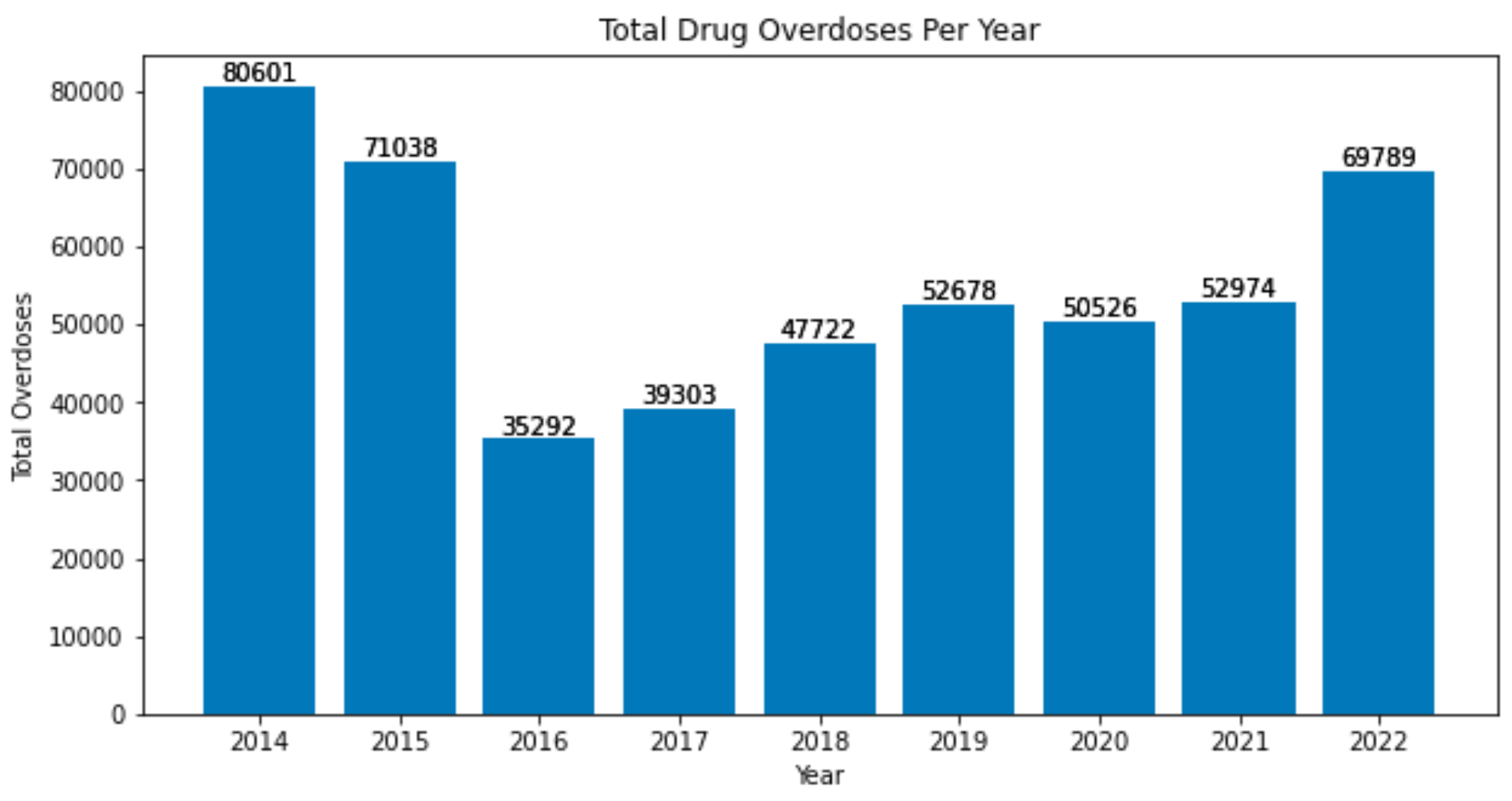
Problem Area

- There is a major opioid crisis in the United States
- Addicts can attempt to seek help in many ways. Social media being one of them
- Can we use social media/ public forums as a sensor for Opioid deaths in the U.S.?
- If we can prove this hypothesis this can give us insights into future deaths from Opioids in the future.

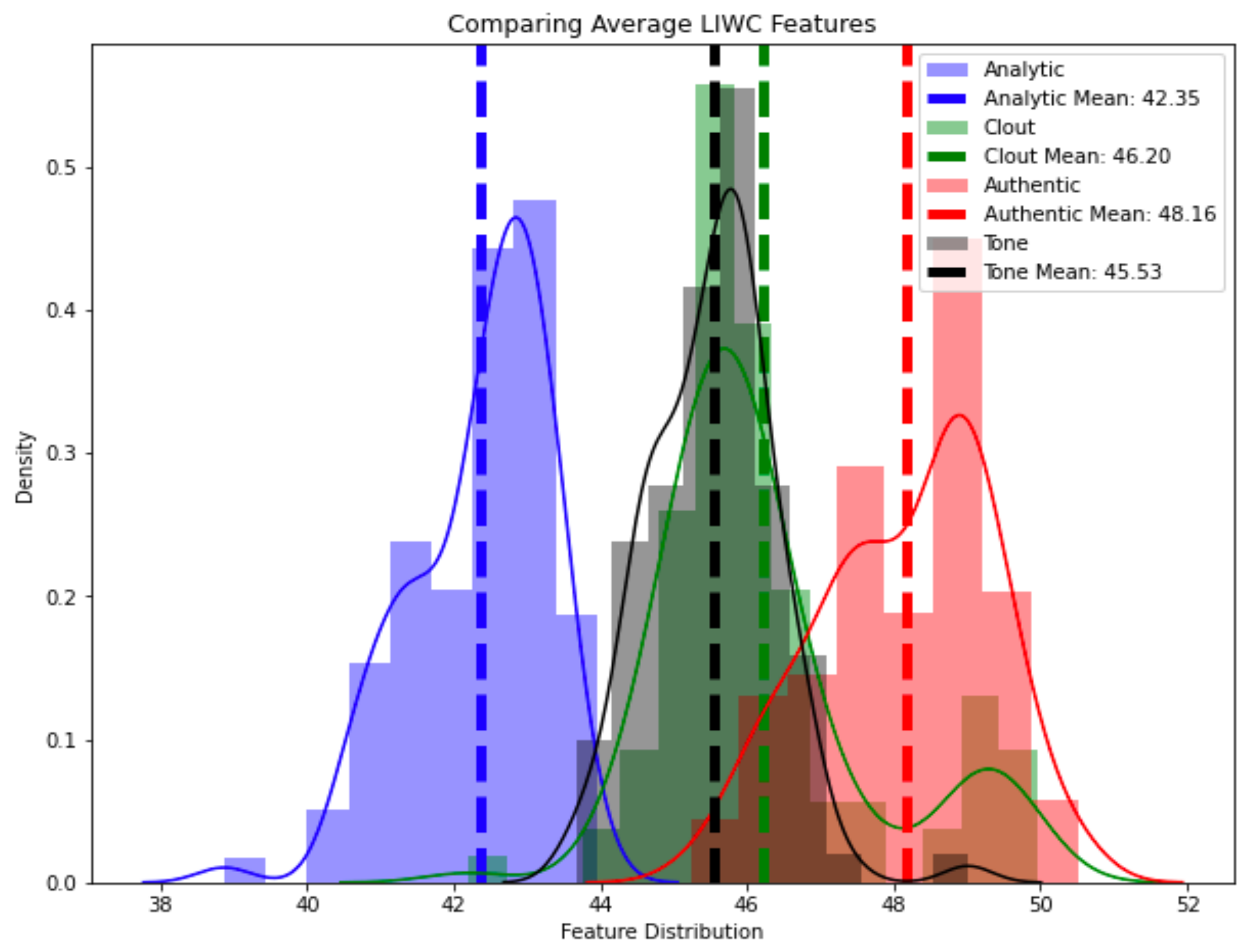
The Data

1. Data contain information about Reddit posts. Using LIWC-22 we can gain insight from the text
2. With over 100 built-in dictionaries LIWC-22 provide insights into their psychological states, including their emotions, thinking styles, and social concerns
3. LIWC reads a given text and compares each word in the text to the list of dictionary words and calculates the percentage of total words in the text that match each of the dictionary categories.
4. Overdose data was found shown in figure below

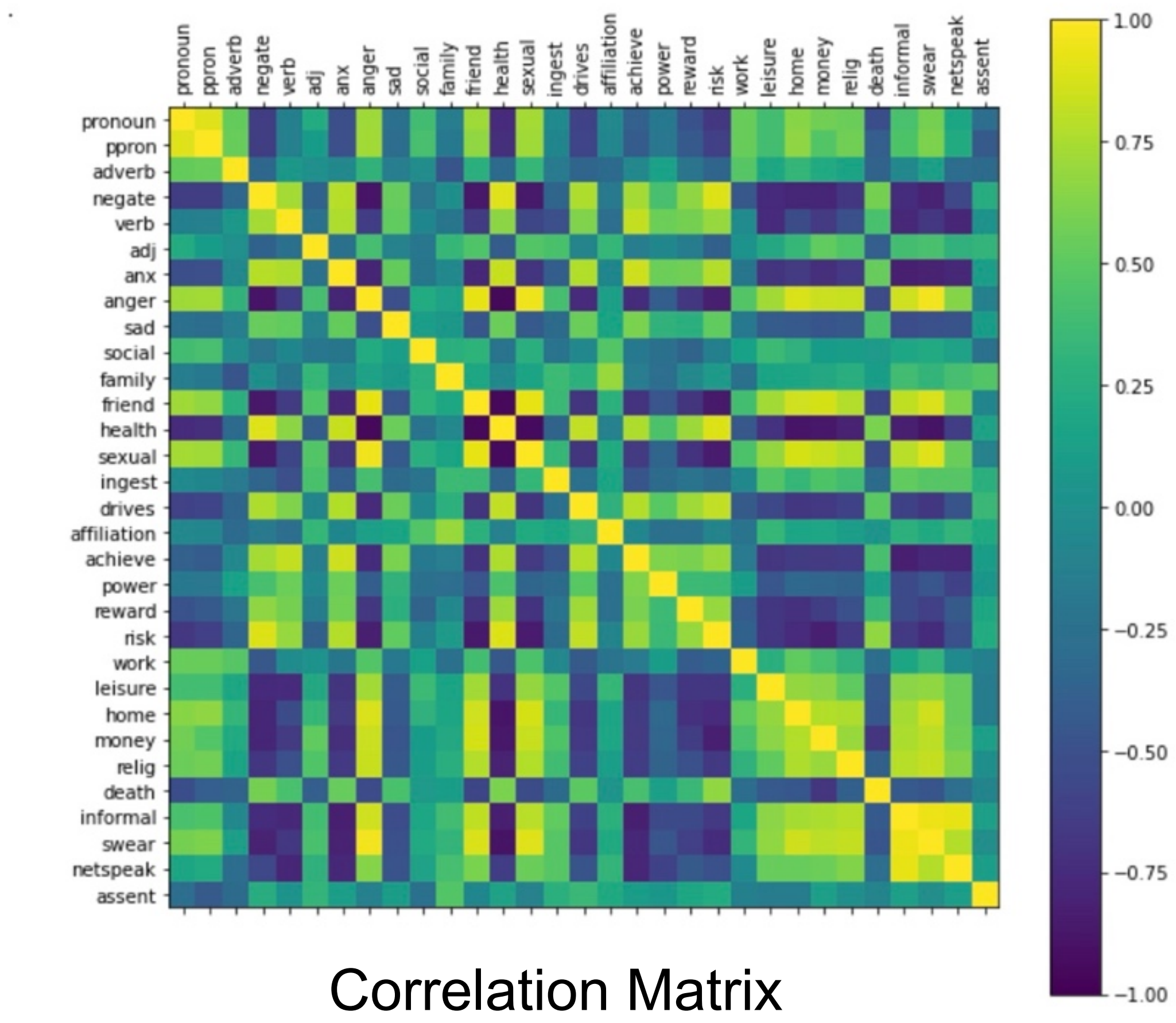
WC	Analytic	Clout	Authentic	Tone
17	7.44	1.00	91.95	1.00
5	93.26	50.00	43.37	99.00
138	76.43	23.44	54.89	78.80
5	52.71	97.69	1.00	99.00



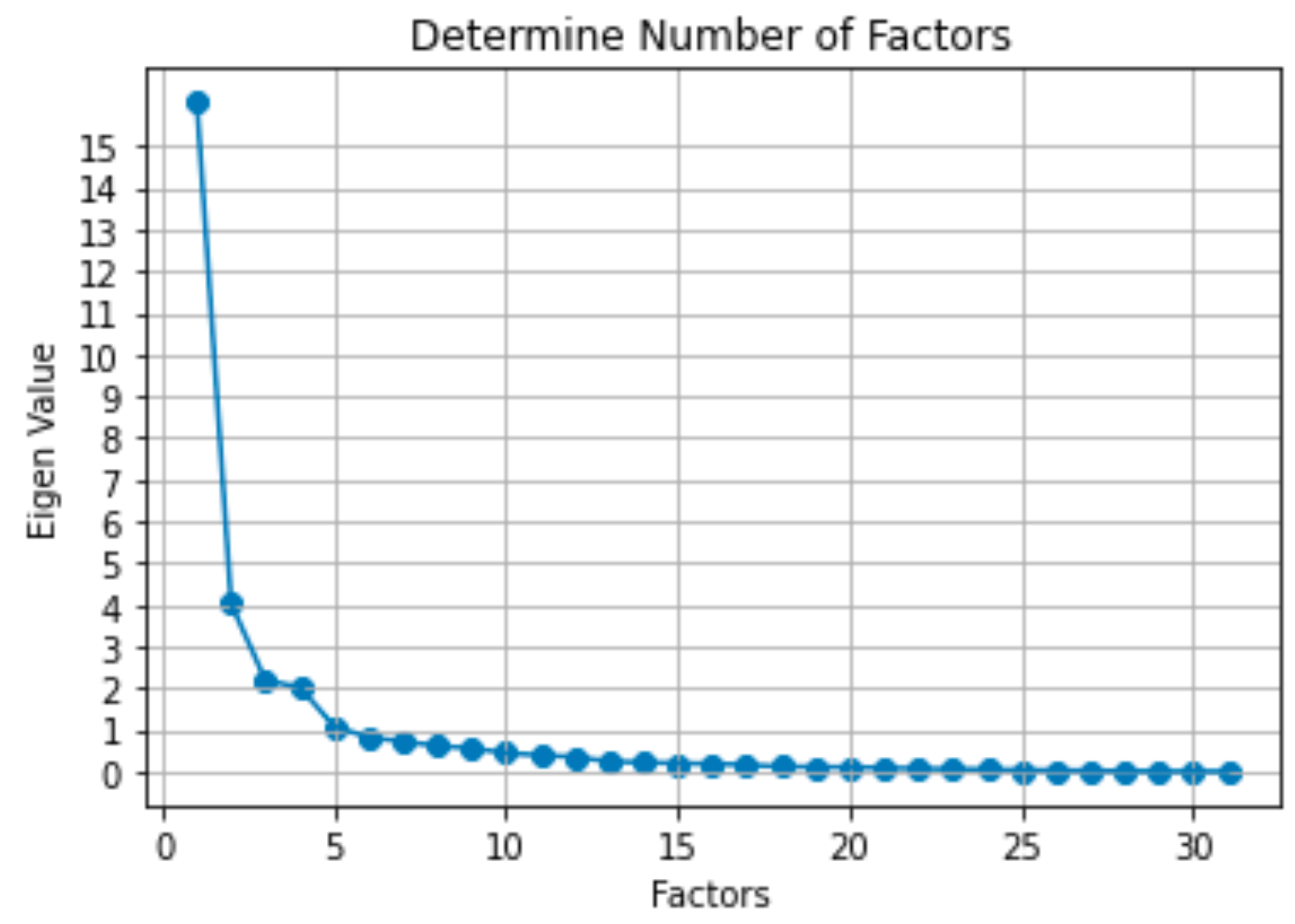
EDA and Statistical Analysis



Distribution of the average LIWC values



Correlation Matrix



Determining Number of Factors

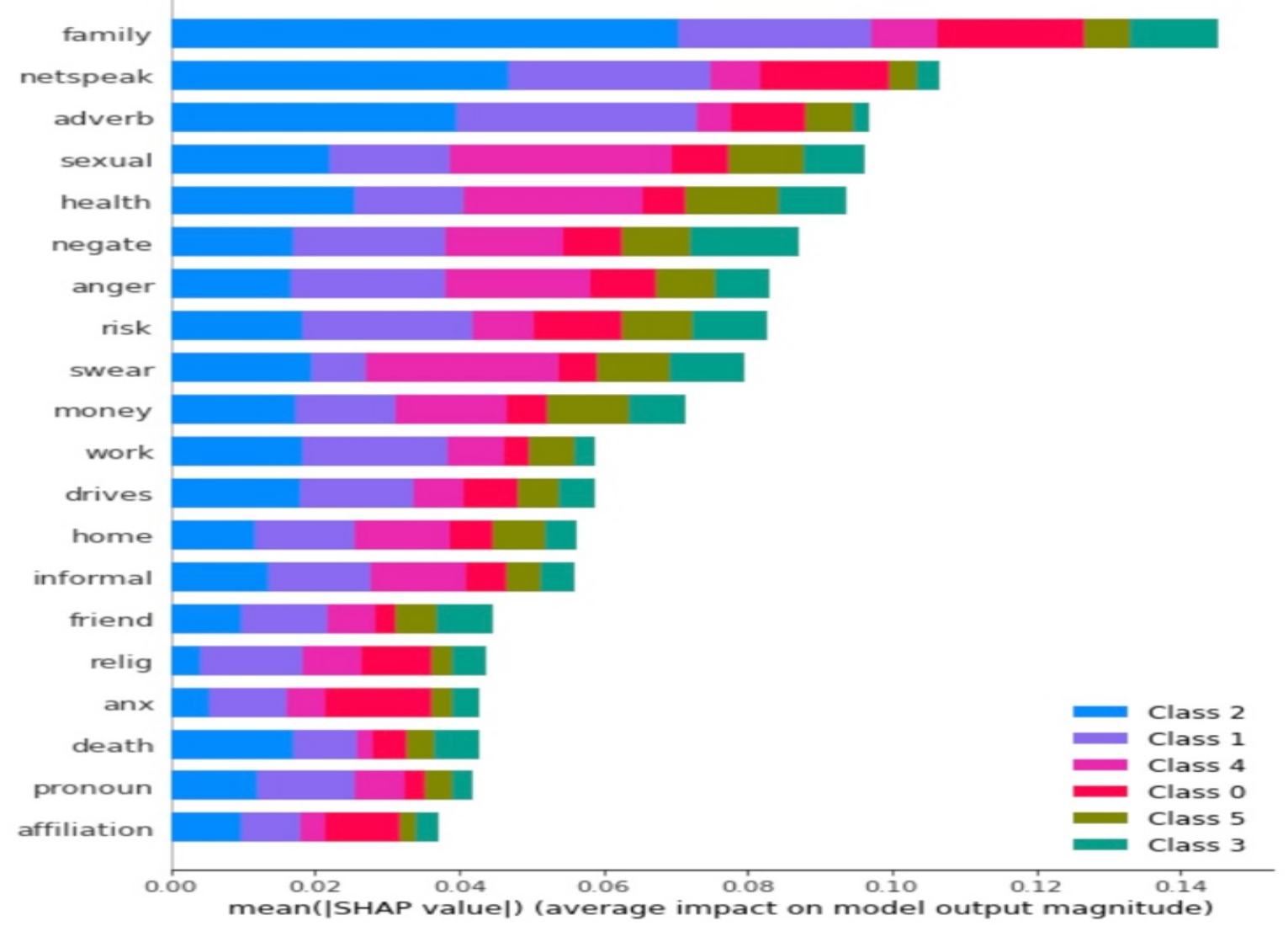
	0	1	2	3	4
netspeak	0.884776	0.040117	0.224192	0.194173	0.195038
informal	0.839116	0.345167	0.307697	0.169830	0.170645
swear	0.720238	0.567439	0.268213	0.152205	0.139603
leisure	0.632404	0.368118	0.010735	0.280154	-0.370591
anger	0.631196	0.723997	0.182655	0.086129	0.013798
friend	0.591599	0.701735	0.202715	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

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.

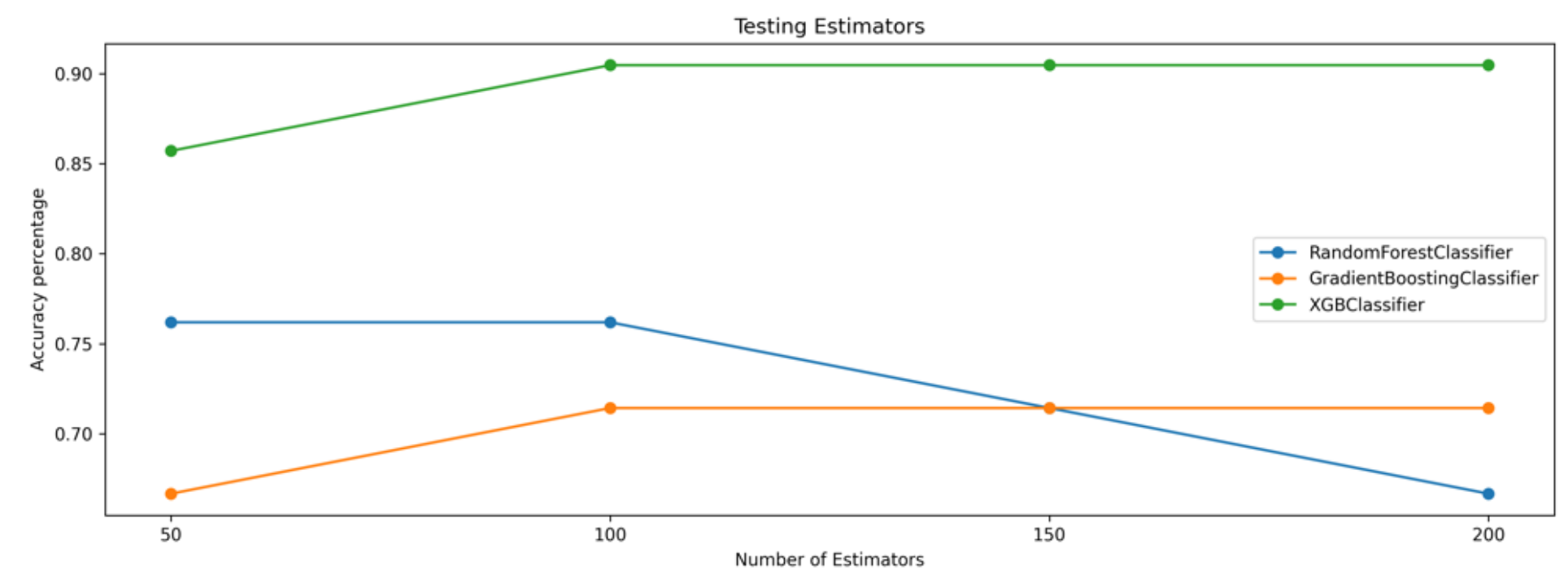
Modeling and Results

Classification

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



SHAP



Determine Number of Estimators

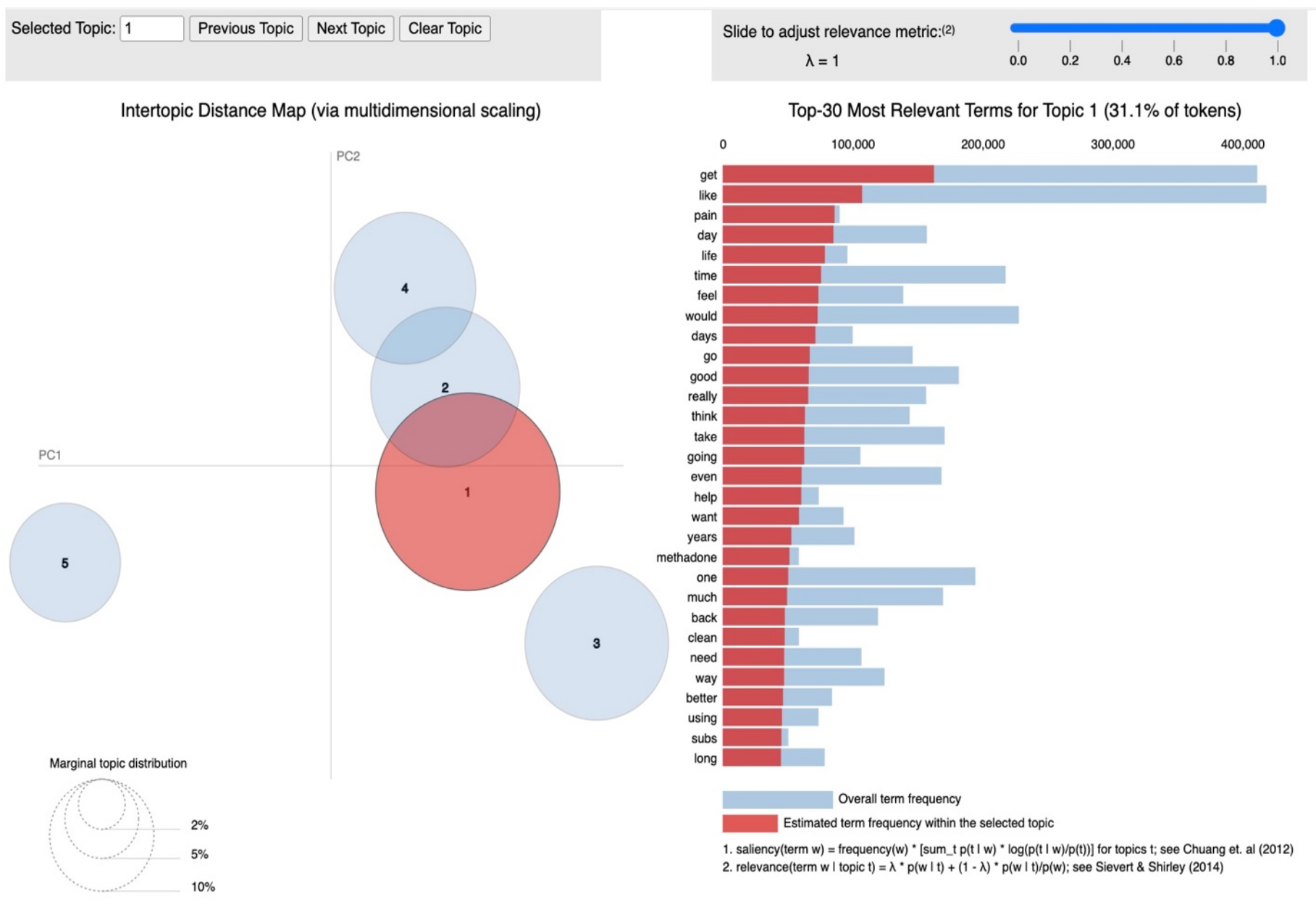
- The SHAP values allow us to understand the contribution of each input feature in predicting a certain class, compared to a baseline prediction.

Multivariate Regression

Metrics	LinearRegression()	RidgeCV(alphas=array([0.1, 1. , 10.]), cv=5)
R^2 Score	0.9154843531345928	0.8948055999265808
MSE	147224.62555340317	183246.61450899945
Cross Val MSE Avg	330155.2250258535	285421.7931919859

LassoCV(cv=5)	XGBRegressor()
0.907290343200958	0.9163068073231104
161498.43270039663	145791.92623167663
323124.52273181203	264124.45000954776

Topic Clustering



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Conclusion

We performed statistical analysis and utilized three machine learning techniques, including classification, multivariate regression, and topic clustering. Our results showed positive outcomes for classification and topic clustering, while regression requires further tuning and data collection. The analysis can model the sentiment of a given forum, and we suggest that these Reddit forums can be used as a sensor for opioid-related deaths. However, further work is needed in feature engineering, model tuning, and data collection. We acknowledge that our data only spans from 2014 to 2022, and accessing a larger range of data could lead to more successful results. In conclusion, our project highlights the importance of choosing appropriate modeling techniques and evaluation methods for successful analysis.

Glossary:
Python – A programming language, capable of processing data/statistical analysis
Pandas – An useful data manipulation package in python
Df, dataframe – Data manipulation structure in R & python pandas
SHAP- SHapley Additive exPlanations
LIWC - Linguistic Inquiry and Word Count

Resources:
Python pandas visualization: https://pandas.pydata.org/pandas-docs/stable/visualization.html#visualization-hist
Scikit Learn - https://scikit-learn.org/stable/
Reddit data - https://www.reddit.com/user/opiates
Overdose data - https://catalog.data.gov/dataset/?tags=drug-overdose&amp;res_format=CSV