Project Deliverable 3 (v1 Final Report)

Michalina Jadick, Della Lin, Jayden Font, Divya Gowravaram, Nikita Jakkam

---

All data should have been collected. All project questions should have been reviewed, answered, and submitted in a written document outlining findings as a PR. You will also be asked to submit the associated data and a README explaining what each label/feature in your dataset represents. Your team should meet with the client before this deliverable.

Checklist

1. All data is collected
    a. Raw data from the web scraper has been fully merged, further processed, (using ‘./data/merge_filter_csvs.py’ and ‘./data/processing_sentiment_local.ipynb’) and the most processed data is represented by ‘GFM_Data_VADER_Sentiment.csv’ in the data folder
    b. ‘GFM_Data_Final.xlsx’ is the full excel file with campaigns extracted organized by keyword which we shared with our client, located in data folder
2. Refine the preliminary analysis of the data performed in PD1&2
    a. Visualizations of the data which play an important role in the data analysis and attribute generation are shown by the figures we have created and are described under our answers to the key questions.
3. Answer another key question
    a. First question: What do we use to define a successful campaign and why? (see below)
    b. Second question: What are some of the trends we see in campaign success over time and what are some techniques we want to use to analyze the data? (see below)
    c. Third question: How is financial need for stigmatized conditions (ranging from rehabilitation services to memorial/funeral costs) framed, and how does that vary across time and by population? (see below)
4. Attempt to answer overarching project question
    a. How is drug use framed-- a moral failing (deviant), or a medical condition in need of treatment (deservingness)? Does the framing affect success? (addressed in bold below)
5. Create a draft of your final report
    a. The key results and data analysis to be addressed in our final report are drafted in our answers to the key questions, including some of the key figures and what they demonstrate regarding the overall project. We plan to organize this more officially in Deliverable 4.
6. Refine project scope and list of limitations with data and potential risks of achieving project goal
    a. See highlights in limitations section in updated project proposal below.
7. Submit a PR with the above report and modifications to original proposal

Overarching Project Question:

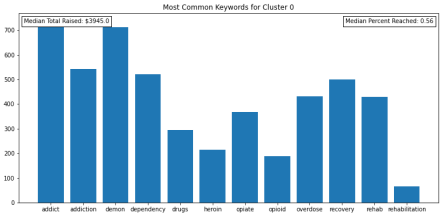
How is drug use framed-- a moral failing (deviance), or a medical condition in need of treatment (deservingness)? Does the framing effect success?

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a sentiment analysis model that produces a score pertaining to how neutral, positive, or negative text is and the "intensity" of that sentiment. VADER scores on a scale of -4 to 4 with -4 being text that is extremely negative and 4 being text that is extremely positive. The model is based on a dictionary that maps words to either positive, negative, or neutral sentiments and changes the intensity score based on the word's relation to the text. VADER is able to pick up on features of text that are commonly found in social media, including Western emoticons and capitalization, and reflect that in its intensity scores. For this reason, we gravitated towards using this model as it is lightweight compared to other NLP techniques and we believed that GoFundMe descriptions would match the social media baseline of VADER as GoFundMe campaigns, like social media posts, are commonly shared amongst a user's network.

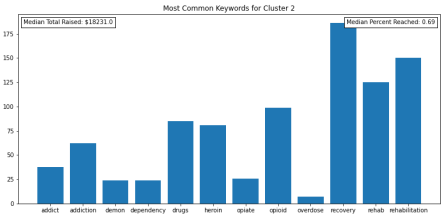
By running sentiment analysis with VADER, we can get a sense of how a campaign is framed, positive or negative, and in turn use that as new attributes to predict success using logistic regression or another model that better predicts based on our data.

Using the sentiment scores from VADER, we were able to cluster the data using a Gaussian Mixture Model to determine if there were any trends that clustering would pick up regarding campaign framing and success. We specifically sought to see if certain clusters contained more successful campaigns based on the total dollar amount raised and what percentage of the goal was reached, and whether or not certain keywords were more frequent in those clusters. When measuring success, the median for each clustering was used rather than the mean to avoid outlier effects. There were 8 clusters total, but two were excluded from this analysis since they only contained either one or two outlier campaigns. The results for the clusters were as follows:

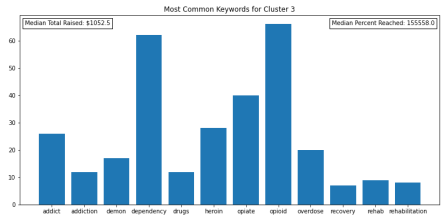
(see ./analysis/Clustering.ipynb)



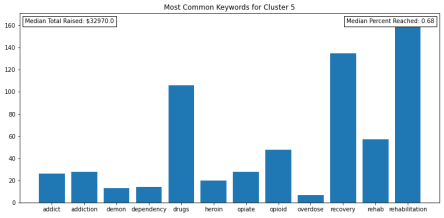
Cluster 0 had a median amount raised of $3945 and a relatively low median percentage reached of 0.56. Because both the total raised and the goal reached were small, it suggests that the campaigns in this cluster (which happened to be the largest cluster by far) were only moderately successful. The most common words for this cluster were "addict", "demon", "addiction", and "dependency". Relative to the other keywords, it is possible to state that these have a more negative sentiment.



Cluster 2 had a large median total raised ($18231) and a relatively high median percentage reached (0.69). Because the total raised is high, the fact that the percent reached is high as well is indicative of these campaigns being very successful. The keywords most commonly used in this cluster were “recovery”, “rehabilitation”, “rehab”, and “opioid”. With the exception of “opioid”, these keywords can be seen as more positive than some of the others. “Opioid” may be common because it is a descriptive word which likely showed up in successful campaigns because it provides context for the campaign’s purpose. Interestingly, the keywords that were most common in cluster 0 were some of the least common words in this cluster (“overdose” was also uncommon), implying that words that had a more positive connotation performed better while words with a worse connotation performed worse.

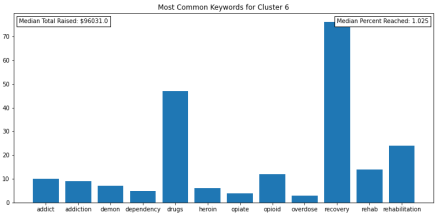


Cluster 3 had a small median total raised ($1052.50) and an unusually high median percentage reached of 155558. Upon further analysis, all of the campaigns in this cluster lacked a goal (either due to not being listed on GoFundMe or being lost when scraping), so it is most likely that these campaigns were grouped together due to the lack of some data skewing the percentage metric.

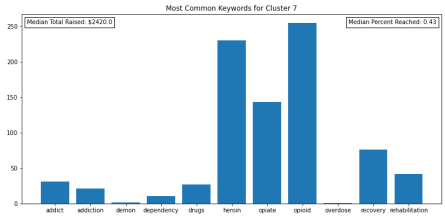


Cluster 5 was very similar to cluster 2. The Median percent raised was 0.68 and the median total raised was ($32970). Like cluster 2, the high amount of money these campaigns brought in and the high percentage of their goal that was reached suggests that these campaigns were

also very successful (perhaps more so than cluster 2). The top keywords in this cluster were “rehabilitation”, “recovery”, “drugs”, and “rehab”. Like cluster 2, these keywords reflect either words with positive sentiment or words that are descriptive (in the case of drugs). Drugs could be considered a negative word depending on the context, however, so its inclusion as one of the most common words is interesting.



Cluster 6 is also similar to clusters 2 and 5, except both the median percentage of their goal that was reached (1.025) and their total amount raised ($96031) are much higher. It is possible that these were the most successful campaigns, although the percent raised being over 1 does suggest that not all of the campaigns had a goal listed. The top keywords are the same as for cluster 5.



Cluster 7 was the worst performing cluster, with a median percentage of 0.43 and a median total raised of just $2420. This suggests that the poor performance was due to campaigns not raising enough to meet their goals rather than missing data (which would reflect in an unusual percentage). The most common words in this cluster were “opioid”, “opiate”, “heroin”, and “recovery”. Other than “recovery”, these words are more descriptive and do not necessarily give off a positive or negative connotation. It is possible that campaigns that lacked language with strong connotation (either positive or negative) and relied more heavily on descriptive terms performed the worst.

To summarize the above results, campaigns that were more successful tended to include more positive sounding words such as “rehabilitation” and “recovery”, while campaigns that used more negative terminology or more descriptive/dry language performed worse. This is in line with our plot showing the percent of the goal that was reached by keyword (see key question 3), which also demonstrated that campaigns with more positive keywords (in general) performed better. Combined, these results all suggest that framing drug addiction as a medical condition that someone heals from (“rehabilitation”, “recovery”, etc.) rather than a moral failure (“demon, “addict”, etc.) did improve campaign success. However, because these relationships were not

reflected 100% percent of the time, the relationships are not directly causational (GMM finds patterns but does not necessarily provide explanations for these results), and missing data may have influenced some results, more in-depth analysis may be needed to confirm these results. Additionally, the manner by which we identified certain keywords as “positive” or “negative” was a combination of sentiment analysis with VADER and human interpretation, meaning the conclusions that we made were partially subjective.

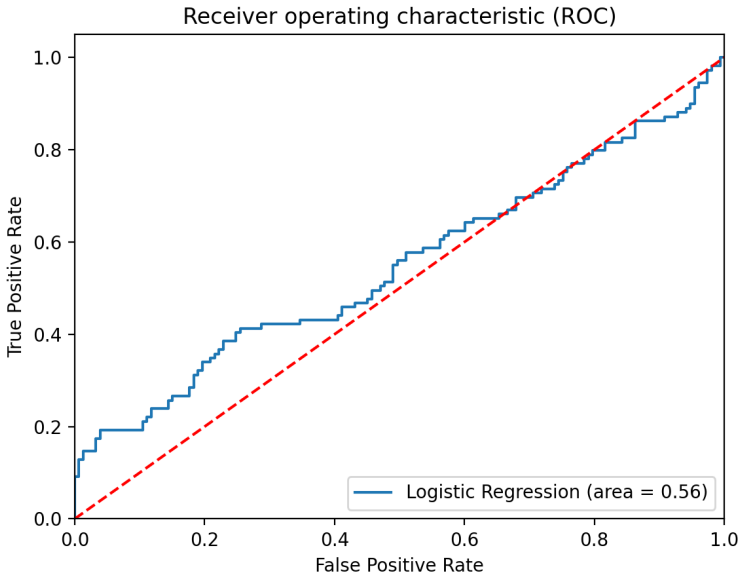
Key Questions:

- **What do we use to define a successful campaign and why?**
  --(from Deliverable 1, updates are highlighted)

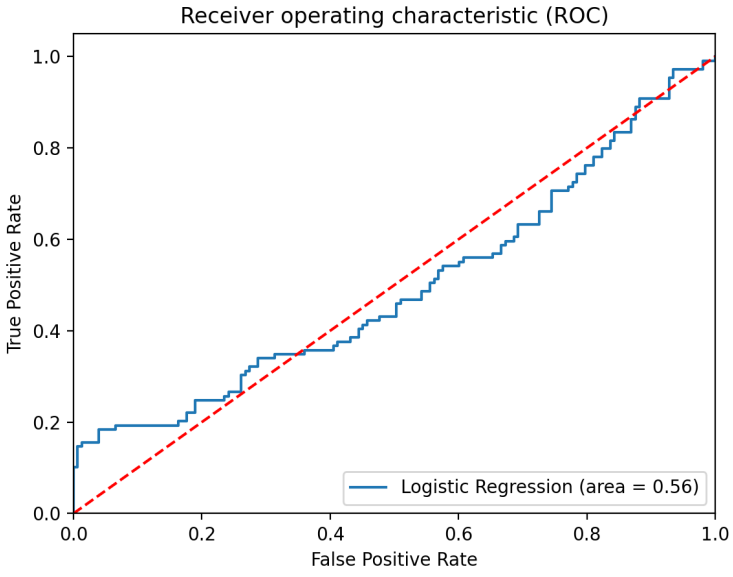
The main metrics we are using to measure campaign success in the preliminary stages are basic quantitative data like the ratio of how much was raised to the campaign goal. We have now obtained this information by implementing our web scraper and intend to move on to uncovering what those numbers represent. Based on the spread of the ratios we obtain, we can set some thresholds that demonstrate success (for example, 80%). We have also measured the number of donors on the page and intend to compare, on average, how much each individual donates to that particular campaign. Another straight-forward metric we have obtained is the number of shares, to gauge success in marketing the campaign. These last two metrics give us a sense of campaign popularity, and we can delve in deeper to determine how correlated popularity is to other attributes like socioeconomic factors, geographic location, etc. Some other factors we discussed are whether campaigns that are backed by charities are more successful than campaigns dedicated to individuals and whether campaigns in certain locations are more popular than others.

To assess how predictable campaign success is based on the set of attributes we have collected for our data, we built a logistic regression model using the quantitative attributes collected, focusing specifically on the year 2017, which saw a spike in campaigns and is known as a time where there was a peak in the opioid crisis. Labels representing campaign success were generated by defining a threshold in the percent reached of the campaign goal, as we described above-- anything above that threshold would be considered successful (encoded: 1) and anything below would be considered not successful (encoded: 0). By building this model, we were able to tune the input definition of a ‘success threshold’ to more directly answer what an appropriate value would be and saw a threshold that results in better model performance is more like 70% compared to the 80% we had previously theorized. The accuracy of the model now is about 63%, which indicates that there is some predictive power behind the attributes we have extracted, which include scores that rank the descriptions and titles of campaigns as positive, neutral, or negative according to sentiment analysis performed (as described under our overarching project question, which addresses sentiment analysis more directly), but there is room for improvement. Principal Component Analysis (PCA) was used to reduce the dimensions of the input attributes and address the issue of high correlation between different inputs, although this did not have a significant effect on the overall accuracy of the model (see ROC curves below). We believe by generating more attributes that relate to the sentiment behind a campaign (for example, assigning weighted scores to each key word used as another indication of how positive or negative a campaign is framed) we can further improve the model.

**See ./analysis/logistic_regression.py



ROC curve for the results of the logistic regression model using compressed attributes (PCA resulted in 4 compressed attributes). Accuracy: 0.63, and there are two attributes with high positive coefficients in the final model and two attributes with large negative coefficients in the final model.



ROC curve for the results of the logistic regression model using uncompressed attributes (no PCA). Accuracy: 0.63, and the attributes with greatest coefficients in the final model are the

positive, neutral, and negative sentiment analysis scores for the campaign titles and descriptions.

- **What are some of the trends we see in campaign success over time and what are some techniques we want to use to analyze the data?**
  --(from Deliverable 2, updates are highlighted)

Some of the trends we are looking to uncover are whether the number of donors, shares, followers have increased or decreased overall (an indication of support to those struggling with opioid addiction), whether there have been more or fewer mentions of keywords in campaign descriptions, and whether there has been an increase or decrease in the number of campaigns under each of the keywords over time. One analytical strategy that may be useful toward this goal is multivariate linear regression, which we can use to uncover these trends inputting multiple attributes of our dataset. We plan to use pandas to convert the data extracted during the scraping phase in our csv files into dataframes for more efficient computation and organization of each attribute column then implement the regression using sklearn. Clustering and classification-based techniques are other approaches we may take for analyzing the data. Performing clustering (with K-Means or a Gaussian Mixture Model) may be a useful first step to identify patterns in the dataset that our own manual analysis would not uncover. Classification techniques such as K-Nearest Neighbors or Logistic Regression may be useful to classify different campaigns as successful (depending on which metrics are used with these algorithms to define “success”) or unsuccessful.

Based on our discussions with the client, we are also looking for specific years that may have had a spike in the number of campaigns and try to understand how current events could have affected that. Using clustering techniques, we may be able to observe what years are associated with higher/lower attributes. We hypothesize that in the late 2010s (~2017), we will observe a “spike” in volume of campaigns and engagement with campaigns since Heather mentioned overdose levels peaked around this time period, and more people may have been seeking support. Additionally, we expect that more current dates will demonstrate a “spike” as well due to another flare in the opioid crisis concurrent with the COVID-19 pandemic.

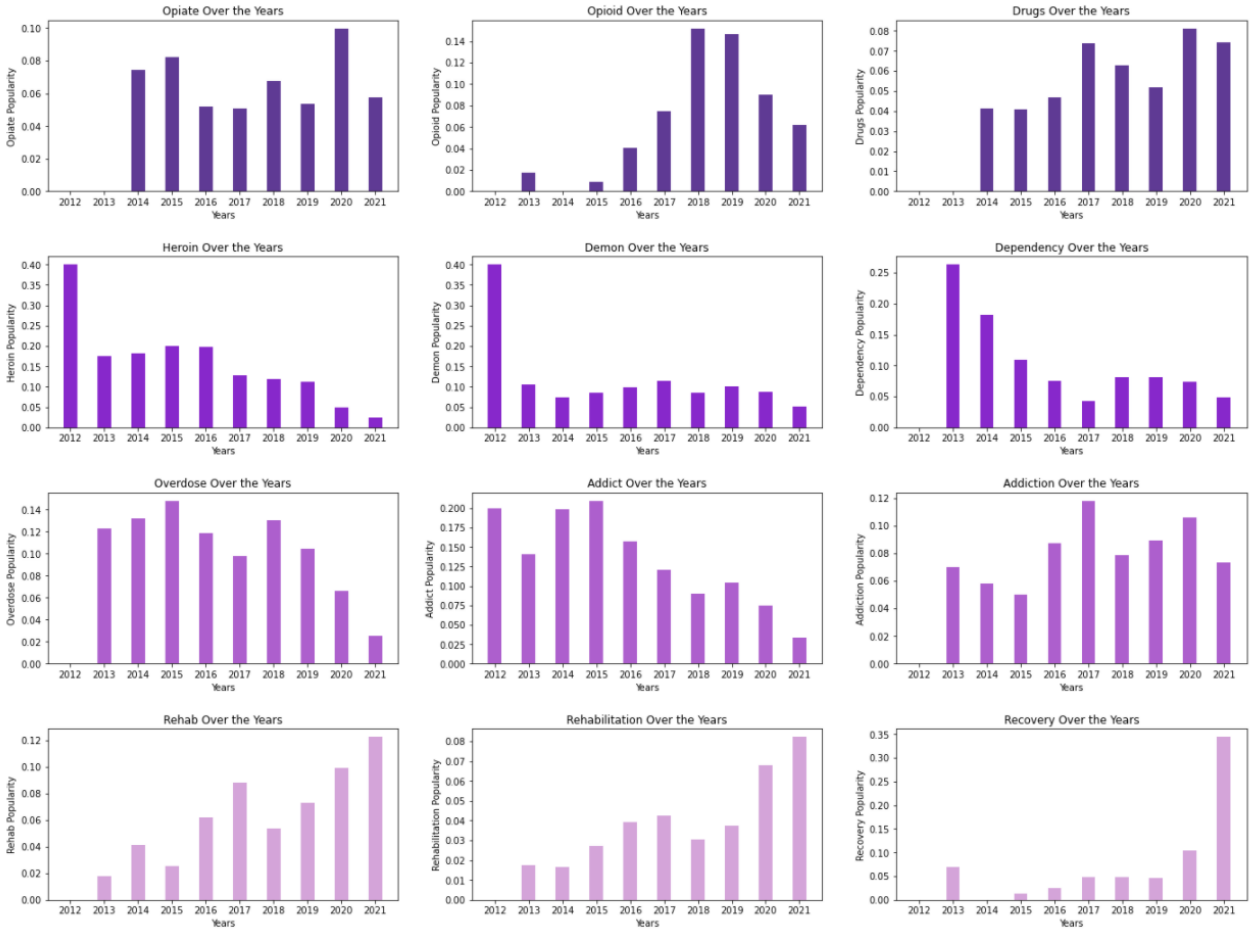
A key aspect of answering this question is to evaluate how change in sentiments is related to campaign success and is variable with time. We intend to analyze campaign titles with an idea of which keywords are more coded vs. overt to see if certain keywords correlate with other attributes and are more or less successful depending on the strength of that language. How these scores affect the predictability of campaign success is addressed in the logistic regression model described above. Then, we can use sentiment analysis to look at how the connotation of campaign titles have changed over time, expecting to see more overt language crop up during times where the opioid crisis was heightened overall. It is uncertain whether the overall trend will demonstrate a tendency towards more overt or covert language over time as the opioid crisis persists.

After getting all of our data, we looked for trends describing the popularity of keywords over the years. Because the number of campaigns differ each year, we normalized the number of

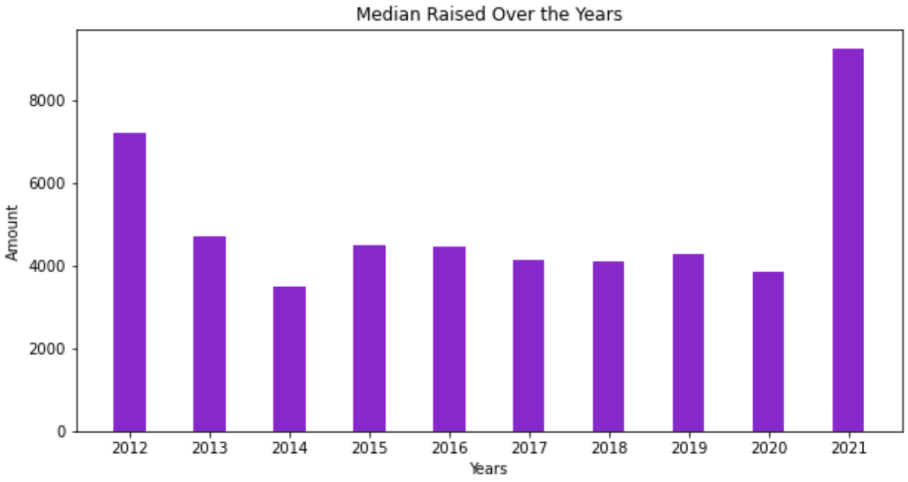
campaigns in relation to each keyword by year. We divided the number of campaigns associated with a keyword by the total number of campaigns that year.

Based on the results, we found that harsher keywords associated with drug use (for example, ‘overdose’, ‘addict’, ‘demon’,etc.) were described more in campaigns in the early and mid 2010s. In contrast, words associated with drug use recovery (for example, ‘rehab’ and ‘recovery’) were much more popular in the late 2010s and early 2020s. There was also a peak in 2017 when it comes to the keyword opioid, which follows our client’s prediction since that year was the height of the opioid epidemic.

(See Below, graph generated from ./analysis/figures.ipynb)



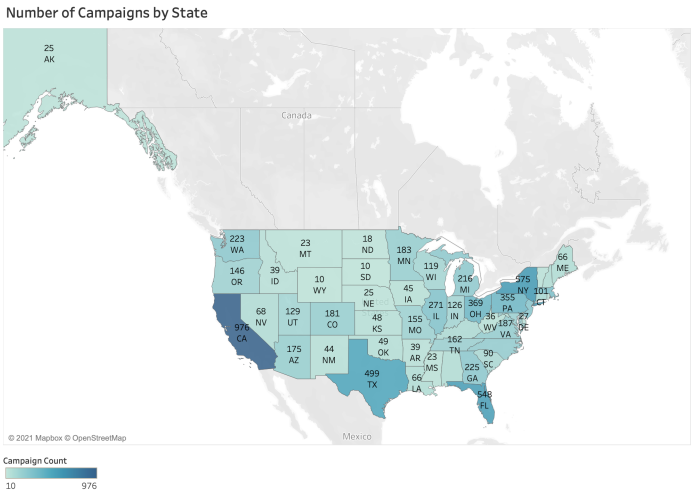
Additionally, we looked at the average amount of money raised each year. We used the median method of calculating the average to avoid severely skewed results from outliers.
(See Below, graph generated from ./analysis/figures.ipynb)

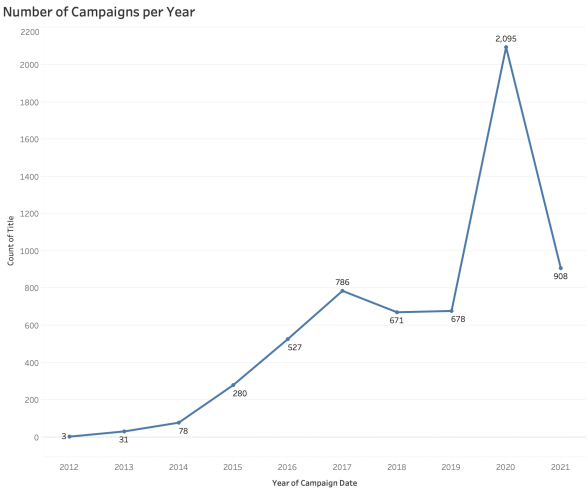


These results correlate to trends we found in other analyses. The peak in 2021 corresponds to the high number of campaigns related to the keywords “recovery,” “rehab,” and “rehabilitation,” which were keywords found in clusters with high percentages of success rates.

- **How is financial need for stigmatized conditions (from rehabilitation services to memorial/funeral costs) framed, and how does that vary across time and by population?**
  --(new key question for Deliverable 3)

Based on Campaigns by State figures, we see a higher concentration of campaigns in more populous states (CA, NY, FL, TX). Variation in numbers can be explained by population density of urban areas in said states. This may indicate how financial need for opioid treatment, whether classified as deviant or deserving, is deemed as more necessary in more urban areas. This idea is further reinforced in the Map of All Campaigns figure, as the campaign density is heavier in the coastal US. Also, in the Campaigns by Year figure, we see the total number of campaigns per year spike in 2017 and 2020. The client mentioned that the US opioid crisis peaked in 2017, but the global spike in 2020 may be attributed to the coronavirus pandemic and its effect on opioid users.



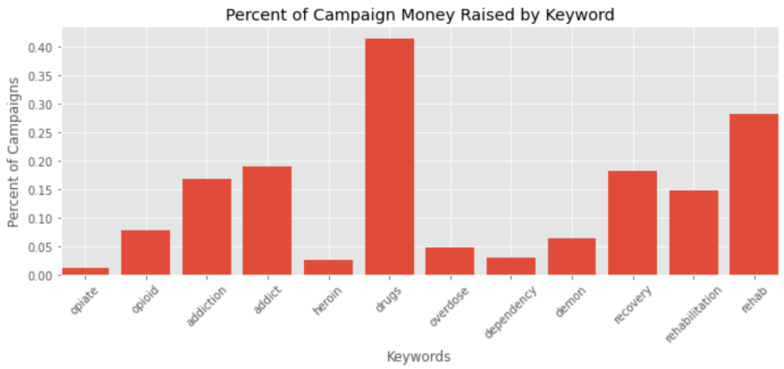


(See Above, graph generated using Tableau)

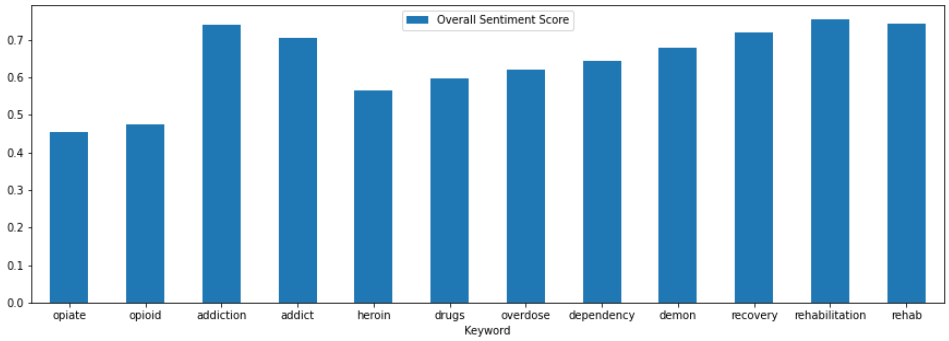
(See Above, graph generated using Tableau)

Based on the figure that shows campaign money raised per keyword, we can see that campaigns that pertain to more positive keywords (like rehabilitation and recovery) or refer to opioids in general terms (like drugs) are more successful at raising funds when compared to campaigns that pertain to more negative keywords (like demon and overdose) or refer to opioid addiction explicitly. The keyword “rehabilitation” is particularly striking as it has the least amount of campaigns associated with it but has roughly the same amount of campaign success (0.15% of money raised) as the keyword “addiction” (0.16% of money raised) despite the fact that addiction had the highest amount of campaigns associate with it. One of the ways we tried to figure out which words were more “positive” or “negative” was to use an NLP model (VADER) for sentiment analysis where keywords with descriptions with higher scores had more positive sentiments than others. Keywords like “rehabilitation” and “rehab” had more positive

sentiment scores than words like “opioid” and “opiate”, but it is important to consider that the number of campaigns associated with a keyword could affect these sentiment analysis scores.



(See Above, graph generated from ./analysis/percent_campaign_money_by_keyword.ipynb)



(See Above, graph generated from ./data/processing_sentiment_locale.ipynb)

Proposal up to date with latest decisions:

****Updates/changes are highlighted below**

Deviance or Deservingness? Opioids, Morality, and Economic Precarity	
Contact	Heather Mooney hmooney@bu.edu
Organization	Boston University - Sociology Department
Organization Description	Sociology, broadly, looks at how macro-level systems, institutions, and ideologies shape life outcomes for individuals and groups. In this project, I explore how the ongoing opioid crisis, which has killed over 760,000 people since 1999, impacts support people and care workers.
Project Type	Data Science
Project Description	For this portion of the research, I analyze crowdfunded campaigns hosted on GoFundMe posted from 2010-2020. Using a variety of keywords, I explore campaigns related to drug-use and overdose to explore how competing frames of drug use and addiction change over time. In addition to exploring how race and gender impact framing and campaign success, I also explore the different relational, moral, and affective appeals that are made to potential donors online. An estimated 10.3 million people aged 12 or older misused opioids in 2017. 9.9 million people misused prescription pain relievers and 808,000 people used heroin. I hope this will be available to them and their networks, amplifying the impact significantly. This project has both policy implications and theoretical promise. Given the long-reaching effects of COVID-19 and the ongoing opioid crisis (which has been overshadowed by and accelerating since COVID-19 began), it will be important to understand how death, loss, and need are constructed by supporting people in times of (layered) crisis. This research represents a case to explore how morality and deservingness change over time and across populations. More broadly, my dissertation explores the “intersections” of social control and "rehabilitative poverty governance." This project provides concrete benefits by centralizing the experiences of care workers and support people--who are often on the "front lines" of service delivery--in order to further improve existing recommendations, policy, and programming.
Data Sets	N/A - I have been hand coding so I can share that, but not sure how useful it will be. GoFundMe data - campaign keyword search. Examples of keywords include: **Opiate** **Opioid** **Addiction** **Addict** **Heroin** Pain medication Pain medicine Pain killer(s)

	Drugs **Overdose** **Dependency** **Demon** **Recovery** **Rehabilitation** **Rehab** **Fentanyl** Unexpectedly Suddenly Epidemic Battle War ***focusing on the words in bold first, to narrow our project scope**
Suggested Steps	1. Scrape data from GoFundMe from 2010 - 2020 using a variety of keywords **with a strategy to optimize the number of relevant campaigns we extract data for** - There may be limitations to getting all data from 2010 to 2020, so a potential alternative we thought of is to break down the years into beginning (2010-2012) - middle (2015-17) - present (2019-2020) 2. Max 300 campaigns per year, all United States postings 3. Include photo data, campaign information, wall posting, and photo information 4. Include social media tagging and relevant pages 5. Clean data (filter out international & repeated campaign) 6. Devise a data visualization tactic to illustrate patterns & findings
Questions to be answered in Analysis	I'm trying to understand how a contested social phenomenon - drug use - is framed as deviant (moral failing) or deserving (medical condition) to a wide audience, and how that **stigma changes over time**. Particularly, I am interested in how this paradoxical problem is framed in relation to need in times of economic precarity and minimal financial/institutional support. Is drug use a criminal act in need of control, or is it a medicalized condition in need of care? What is construed as deviancy versus deserving of support? How is financial need for stigmatized conditions (ranging from rehabilitation services to memorial/funeral costs) framed, and how does that vary across time and by population? ● How is success of a campaign determined by humanizing descriptions/components, race/status of the victim and who is writing the campaign, social media interaction, etc.? ● What do we use to define a successful campaign and why?
Additional Information	**Tools and Methods** For scraping - Scrapy and Selenium webdriver (for searching GoFundMe and finding relevant campaigns), and Beautiful Soup (for scraping actual campaign data). ● Potential reference: https://github.com/lmeninato/GoFundMe ○ We will use this reference as a template, but it makes more sense to build a scraper from scratch so we can make it fit our needs ○ Jane also attached resource: https://github.com/automaticalldramatic/vue-node-scraper For cleaning and preprocessing use Pandas to organize the dataset into dataframes for faster computation.

	Data visualization libraries such as Matplotlib, Seaborn, and Bokeh (interactive web-integratable visualizations).
Limitations / Potential Risks	<u>DATA COLLECTION STAGE</u> ● Since this a new project, we would have to scrape data ourselves before data analysis could be done which means that data analysis would have to take place later in the semester than expected ● Because Selenium takes over the user's computer, it may require a lot of time to gather data which could prevent the owner of the computer from using their device for prolonged periods of time. In this case, we may need to look into using another device or remote access ○ Jane- unfortunately, might not be a solution to this. Check back if we run into problems later on. ○ If we don't mass scrape too much, shouldn't be an issue. ● GoFundMe does not appear to have a way of filtering results by year, so it may prove difficult to get data from GoFundMe campaigns that go all the way back to 2010 (possible solution: we could potentially get all the campaigns and then sort by year after) ● How will we be handling images? Even downloading all the images would be a very heavy task. Will we focus on numerical data like the number of photos on a page? ○ We've decided to focus more on easily-extractable data for now and leave more complicated tasks for the Heather's "deep dive" stage ● Fetching comments may be difficult as it requires Selenium, and the number of comments varies for every page (this makes it hard to write a script to gather this information). Number of comments is much easier to get (could tie into engagement as a metric for success). ● Some search results that appear in later pages return "Campaign Not Found", most likely meaning they were deleted but are still showing in searches. We would need some way to remove these posts? ● Campaigns which are not currently accepting donations may not have the campaign goal listed — We could use a Try/Except Block to handle this without running into major issues ○ Later search results may lack data fields, meaning we may have empty fields for some campaigns ● Need to make sure that we don't have too many duplicate campaigns-- add an if statement to exclude potential duplicates ● When running the web scraper, we have been encountering 403 errors which either break the code or result in gaps in the data with empty campaign results (if bypassed with Try/Except blocks). This may be because the GoFundMe website is blocking suspicious activity and not allowing access to some webpages. We have also encountered ConnectionResetError/ChunkedEncodingError when running the scraper, which may have something to do with our internet connection. <u>ANALYTICAL STAGE:</u> ● Without extracting full descriptions from campaigns (which we determined is very time expensive), we are limited in how much content we have to analyze with natural language processing. The focus for now is campaign titles, but we are not yet sure if this will provide enough sentiment analysis to be analytically significant.

	○ Plan is to start with the titles, and if we run into stall points with the analysis return to this later ○ We ended up implementing the extraction of full campaign descriptions to use for sentiment analysis ● It is difficult to translate quantitative data to complex questions about the socioeconomic and demographic relations to opioid use ● Logistic Regression Model: ○ At this stage, we are unsure how the performance of the logistic model that predicts campaign success can be improved by adding more attributes related to sentiment analysis. It seems the model has a really high sensitivity but a low specificity and hope this can be improved. ○ The downside to using PCA is that we are not sure from which original attributes the compressed features come from-- because of this, it's difficult to tell how the original attributes we extracted are grouped together and impact the model. We may need to try reducing dimensions to fix the issue with colinearity of features using another method.
Questions for Client	1. Differences between content analysis & selective discourse analysis? 2. We were able to generate sentiment scores for each keyword, and we wanted to ask if the results were similar to the client's list of rankings of the keywords?