Project Deliverable 2

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

----

Sufficient data should have been collected to perform a preliminary analysis of the data and attempt to answer one question relevant to your project proposal which you will submit as a pull request. If data has already been collected for your project you must answer two questions.

Checklist-- Overview

Collect and pre-process a secondary batch of data

- We currently have a scraper that extracts all of the URLs of the pages containing the words from a list of search words we specified ('gofundme_scrape.py')- The data folder of the repo contains 'urls.csv' which organizes urls extracted with the list of keywords that were found therein.
- We created a scraper to extract the following information from each campaign ('campaign_info_bs4.py'):
    - Title
    - Organizer
    - Beneficiary
    - Date created
    - Location
    - Campaign tags
    - Number of followers
    - Number of donors
    - Number of shares
    - Amount raised compared to the total goal (as a percentage)
    - Number of campaign updates
    - Number of campaign comments
    - Country code
    - Currency code
    - Number of donations
    - Is a beneficiary a charity? (True/False)
    - Beneficiary charity name (if given)
    - Is a beneficiary a business? (True/False)
    - Are campaign organizer(s) part of a team? (True/False)
    - Does a campaign have comments enabled (True/False)
    - Does a campaign currently have donations enabled (True/False)
- In terms of preprocessing, we also added functionality to our scraper to detect missing information, merge duplicate URLs by keyword, and format excel file ('data_formatting.py').
- An excel file compiling the most comprehensive amount of data we have collected so far (with sheets organized by keyword) using an earlier version of the 'campaign_info_bs4.py' scraper is in in the data folder of the repo

(‘GFM_Data_Test.xlsx’). The final step is running our updated scripts in batches between team members to avoid errors, capitalize on time, and obtain a filled version of this dataset.

Refine the preliminary analysis of the data performed in PD1

- Because we did not have a dataset to work through from the start of this project and are building it ourselves, our preliminary analysis of the data focuses on our takeaways from the process of building our data scraper and our hypotheses/strategies for the data analysis stage.

Answer one key question

- First question to address: What do we use to define a successful campaign and why? (see below)
- Second question to address: 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)

Refine project scope and list of limitations with data and potential risks of achieving project goal

- See highlights in limitations section in updated project proposal below.

Submit a PR with the above report and modifications to original proposal

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.

- **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?**
  --(new for Deliverable 2)

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. 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.

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	DATA COLLECTION STAGE ● 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. ANALYTICAL STAGE: ● 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 ● It might be difficult to translate quantitative data to complex questions about the socioeconomic and demographic relations to opioid use
Questions for Client	1. How much data is sufficient to start (i.e., for each time period how much should we collect at first?) a. Since we cannot filter results by year directly on GoFundMe, we will have to determine this retroactively after scraping key term by key term. i. Original proposal wants 2010-2020, there have been several iterations since 80s,90s. ii. Peak = 2017-2019? (2015-2020 will still show a lot of variability) 2. Differences between content analysis & selective discourse analysis? 3. Since many of the GoFundMe campaigns are not explicit when it comes to disclosing opioid addiction, how would we know that campaigns being scraped are associated with opioid/drug addiction and are relevant to the study? a. We won't know exactly, but as long as we get the most information possible -- we can triangulate, and search those people specifically up on social media, obituaries to get more information (we may not be responsible for that) b. Have a column that flags if it is specifically mentioned-- but its ok if we don't get that direct disclosure (indicate how many key words are included in a single campaign) c. Note-- keywords are validated in the literature, so its not unusual 4. Is it possible to reach out to GoFundMe to inquire about gaining access to data related to our project? a. Jane will send an email with us cc'd b. Heather can also email or have her advisor email