COGS9: Introduction to Data Science

Assignment #2: Final Project Proposal

Due date: Friday 2024 November 01 23:59:59

Grading: 10% of overall course grade; 40 points total

Download the editable version of this document and add your responses in the locations indicated. Please respond using the blue font color used in the response text, as it makes the assignments easier to grade. This is assignment is to be completed *as a group*, with only one submission for the entire group. Once completed, save the document as a PDF and submit on Gradescope. Be sure to assign pages to each answer when you submit (see the Gradescope instructions here[1].)

Group Member Information:

Assignment will be completed in a group of 4-5 people. These will be the same people you work with for the final project. This assignment will help get you started on the final project.

First Name	**Last Name**	**PID**
Jaden	**Goelkel**	**A18247795**
Richard	**Wang**	**A18222141**
Yathin	**Mrudul**	**A18194759**

Leo	**Wong**	**A18472480**

Question

Clearly state the specific data science question you're interested in answering. (5 pts)

How does the emotional valence of New York Times headlines over the past 3 decades correlate to the suicide rates in the United States?

Hypothesis

Write down your group's hypothesis to your question. Provide justification how you came to this hypothesis. (What background information or instinct led you to that hypothesis?). (10 pts)

Our hypothesis is that there's a correlation between the titles of the New York Times News in the past 3 decades and the suicide rate in the U.S. which can be used to identify what type of words and phrases the New York Times should prevent using in the future in order to decrease the suicide rate in the U.S.

We believe this hypothesis is justified because, in modern society, media has a profound impact on people's everyday lives. It shapes opinions, influences human behaviors, and can even affect an individual's emotions through exposure to various narratives, images, and stories. The New York Times is one of the most widely read and trusted news sources in the U.S. If headlines or stories are framed with uplifting and positive content, it might make the readers feel more motivated and happy. On the other hand, if the headline includes distressing language or focuses on negative topics, it might contribute to an increase in anxiety and stress among readers. This might be a significant cause of the potential increase/decrease in suicide rate in the U.S.

We can design a model that analyzes the amount of negative content (ie. words, phrases, topics) within New York Times News titles over time and the suicide rate in the U.S. over time years. If there's a correlation, then it can help alert the newsletter editors to not include certain words, phrases, or topics when writing their title in order to not trigger stress and anxiety among the public, thus hopefully decreasing the suicide rate in the future.

Background Information

Include a few paragraphs of background research and information on your topic. This should include at least 2 citations to work from others. Including hyperlinks to reputable sources is fine. (10 pts)

The influence of public media on mental health is well documented. Studies highlight how exposure to negative news can intensify psychological stress. Research shows that this effect is particularly strong with repeated exposure to negative stories. According to the American Psychological Association, "media overload" can lead to increased levels of anxiety, stress, and trauma-like symptoms. This is especially seen as people have near-constant access to news in the digital age. This persistent exposure to distressing headlines creates cumulative mental strain, often contributing to feelings of helplessness or despair (APA, 2022).This phenomenon, combined with the influence of widely read news outlets like the New York Times, may be significant in understanding the impact of media on U.S. suicide rates, which have been a rising concern over recent decades (Centers for Disease Control and Prevention, 2021).

Given the New York Times' substantial role in shaping public discourse, examining the emotional tone of its headlines over time could reveal patterns that align with trends in suicide rates. If such a correlation exists, findings could inform strategies encouraging news organizations to frame content in ways that prioritize public mental health. Such guidelines could help minimize stress triggers for readers, potentially improving mental health outcomes.

https://www.apa.org/monitor/2022/11/strain-media-overload

https://www.cdc.gov/

Data

Include a description of the perfect dataset you would need/want to answer this question. How many observations would you need? What variables would you collect? Explain the perfect dataset that you would want to answer this question.

The perfect dataset would include all of the New York times headlines in the past 3 decades with that has its associated date. All of the data would need to be tokenized to be able to perform a sentiment analysis. The data set would need to be all valid New York Times headlines without any gaps in dates or censorship in the headlines. Additional we would need the data on all the suicides in america over the same time period. This data would need to contain many elements such as the age demographics of the individuals, suicides per 100,000 people, data on a daily timeframe that can then be grouped into weekly, monthly and yearly intervals. The data would need to be accurate and best reflect all of the suicides in the given time frame. Both of the data sets would both be as comprehensive as possible with multiple different dimensions that would allow for further analysis. I would want the data to be organized in 2 dimensional csv that has all the associated data about the headline or suicide on the same line item.

Then, look online for available datasets. Find a dataset that could be used to answer this question. Describe how many variables are included, what those variables are, and how much data has been collected. Discuss the dataset's limitations and how it differs from your ideal dataset. Alternatively, you can collect your own data. You must explain what information you will collect and how you are going to collect this data. These do not have to be collected by the time this proposal is submitted, but they must be collected by the time the final project is submitted. (5 pts)

The first dataset that I found was a dataset with all the new york times data in the last 30 decades. This dataset has the headline title, the source of the headline, the word count of the article, the url of the article, the print section, the author, the type of publication, and the publication date. There are about ~100,000 unique line items so this dataset is very comprehensive with the headlines in the past 30 decades. Some of the biggest limitations with the dataset is quickly and cheaply checking the validity of all the entries but assuming the data is correct this dataset fits most of the needs for the project because we can easily perform an analysis on the headline and the emotional valence associated then use the date of that publication to the correlation that with the suicide data. This data is organized in a csv file that would be easy to use.

The second dataset that I found is an overview of the suicide rates from 1985 to 2016. The dataset has, country, year, sex, age, suicide number, population, suicide per 100k, country-year, and the Gdp per that year. Assuming all teh entries are accurate and all the available data points were recorded this dataset fits the needs to answer the problem because the suicide rates in america can be graphed over time and deeper analysis can be performed. The Ideal dataset would be very similar however the validity of the data would all be verified and the collection method would be transparent. The data is organized in csv files that would be easy to use.

https://www.kaggle.com/datasets/johnbandy/new-york-times-headlines

https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016

Ethical Considerations

Read the data science ethics checklist from lecture. Then, discuss what ethical considerations must be made when answering your specific data science question. Brainstorm and explain how you would address these considerations for each of the following categories in your specific project: Team Bias, Sampling Bias, Data Bias, Consent, Data Privacy / Ownership, Algorithmic Bias / Discrimination, Transparency, Unintended Consequences, Continued Monitoring / Accountability. Feel free to write about additional ethical considerations you would make that aren't included on the checklist. Note that data privacy is NOT the only ethical consideration for a data science project. It is a piece, but there is a lot more that has to be considered. (10 pts)

Team Bias:

It is vital to consider potential implicit bias within the team when approaching such a project. In order to avoid this, we plan to periodically communicate and reflect on our opinions upon the topic with each other, allowing us to come to a common unbiased consensus. We intend to make it certain that each team member justly considers how their perspective may influence the analysis and interpretation of our results.

Sampling Bias:

We ensured that our process of data collection was absent of any sampling bias by utilizing a representative sample of U.S. suicide rates and New York Times headlines. This was done so by using data of suicide over a large period of time(past 3 decades and 1985-2016). An example of possible distortion within the data would be if the dataset disproportionately included headlines that were popular or politically charged. By adamantly verifying that the dataset encompasses a wide array of topics and appropriately represents a diverse sample of news published by The New York Times, we were effectively able to avoid such cases of potential sampling bias.

Data Bias:

As touched on above, our team addressed the issue of data bias by analyzing whether or not certain article headline formats or demographic data on suicide rates are overrepresented or underpresented. For instance, sentiment analysis as a whole may be distorted if certain news categories—like political or crime reporting—are more commonly labeled as being negative. Thus, our team will utilize a unique variety of headline topics and employ a balanced sentiment categorization within our dataset.

Informed Consent:

Due to the fact that we are using publicly accessible data sets rather than personal data,

concerns of informed consent are less significant in this project. However, nonetheless, we continue to adhere to the ethical guidelines of using aggregated datasets that may contain sensitive information by explicitly acknowledging our data sources.

Data Ownership and Privacy:

Although our data is openly public and accessible, we still maintain that it is crucial to take privacy and data ownership into account. With this, we have appropriately credited sources and adhered to usage policies set forth by dataset providers (like Kaggle). In order to protect people's privacy, we have only use aggregated and anonymised data regarding suicide.

Algorithmic Discrimination:

Algorithms used in sentiment analysis may display bias, particularly if they were trained on biased datasets. For example, some terms may be disproportionately linked to negative sentiment, which may not accurately reflect the content of headlines. To lessen this, we will assess how well the sentiment analysis tool performs over a variety of headline subjects to make sure it is accurate and fairly balanced across headline kinds.

Transparency:

We will meticulously record all aspects of our processes, including data pretreatment, modeling decisions, and constraints. Others will be able to comprehend, replicate, and evaluate our analysis as a result. To prevent inaccurate results, we will also provide clarification on any assumptions made, including sentiment classification and data interpretation.

Negative or Unintended Consequences:

Determining whether media sentiment and suicide rates are correlated is a delicate endeavor since it may result in unforeseen conclusions. For example, assuming causation when there is merely association can lead to oversimplified conclusions. To avoid this, we will make sure that our findings are presented with caution, emphasizing that correlations are meant to direct future research rather than draw firm conclusions and do not indicate causation.

Continued Monitoring of Biases:

Due to the possibility that biases can change over time, it's important to review the model on a regular basis to make sure it's still objective and applicable, particularly if it's being utilized in practice. If this study continues into the future, we will plan for frequent assessments and recalibrations of the model to adjust to shifting linguistic or social contexts.

[1] https://guides.gradescope.com/hc/en-us/articles/21864315441677-Submitting-a-PDF-for-an-assignment