COGS9: Introduction to Data Science

Final Project

Due date: 2024 December 12 23:59:59 (Thursday)

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

Completed as a group. One submission per group on Gradescope.

Group Member Information:

Please read the COGS 9 team policies to best understand how to approach group work and to understand what the expectations are of you in COGS 9.

First Name	Last Name	PID
Jaden	Goelkel	A18247795
Richard	Wang	A18222141
Yathin	Mrudul	A18194759
Leo	Wong	A18472480

Question (2 pts)

Clearly state the specific data science question you're interested in answering. This question can be the same as what you submitted for your project proposal. Alternatively, you can edit your original question or change your topic completely.

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 (2 pts)

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?). You should incorporate the feedback you received on your proposal.

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 (3 pts)

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 are fine.

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 (2 pts)

Include a description of the perfect dataset you would need 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.

Then, look online for available datasets. Find a dataset that could be used to answer this question. Describe how many observations are included and what variables have been collected. Discuss the dataset's limitations and how it differs from your ideal dataset. If you collected your own data, explain what information you collected, from whom you collected it, and a link to the data.

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.

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 (3 pts)

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.

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.

Analysis Proposal (15 pts)

Here, you will propose how you would use and analyze data to answer your question(s) of interest. You are neither expected nor encouraged to carry out the analyses to answer your question(s). You will describe, in detail, what you would need to do to prepare your dataset for analysis (data wrangling) and what type of analysis you would do to answer your question(s). Explain which how your proposed methods / approaches would allow you to interpret the results from this analysis. We are looking for the correct conceptual understanding and application of ideas discussed in class, not specific and technical implementations. For example, if you are applying machine learning to some categorical data, it’s important to specify whether you will be performing regression or classification. If you are unsure about the details of anything above, ask on Piazza, come to office hours, and/or do further research on your own (Stack Exchange, Google, Wikipedia, etc.).

Specifically, you are required to incorporate *at least four different methods*, exploring ideas from a combination of:

- Data Collection (web scraping, APIs, etc.)
- Data Wrangling
- Descriptive & Exploratory Data Analysis (summary stats, correlation, etc.)
- Data Visualization
- Statistical Analysis (Inference, A/B testing, etc.)
- Predictive Analysis (machine learning, classification, regression, etc.)

- Text Analysis (Sentiment Analysis, TF-IDF, etc.)
- Geospatial Analysis (choropleth maps, geospatial statistics, etc.)

Data Collection

Our analysis would start by collecting the necessary data, including categorical data like New York Times headlines and quantitative data like suicide rates in the U.S. over time. To do this, we will be using the two datasets proposed in the "data" section (linked below). These two data sets contain sufficient amounts of data with New York Times headlines from 1990 to 2020 and countries' suicide rates from 1985 to 2016 which allows us to perform a reasonable analysis. We will make sure to not rush the data collection process as it's important to perform analysis with high-quality data and wrangling methods in order to produce insightful and meaningful results. Links to datasets:

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

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

Data Wrangling

For the suicide data, we will first filter the data to only contain suicide data from the United States as the dataset contains data from multiple countries. We will be careful not to consider variables such as gender and age as these might introduce possible biases. To achieve this, we can use Python libraries such as NumPy and pandas to combine the total suicides from each gender and age within a year and divide by the total population of each gender/age for that year to find the suicide rate for that year. We want to use proportion in order to generate the data as the population changes every year. We will also remove the unnecessary columns in the suicide dataset such as suicides/100k pop, country-year, HDI for year, and gdp_for_year($). For the New York Times headline, we will need to create a new column for the data set and categorize each headline to be either positive or negative, marking negative with numbers closer to 0 and positive with numbers that approach 1. We can then calculate the proportion of negative New York Times headlines each year and create a new dataset with only the year and Negative headlines scores which is the Compound Score achieved from the model*. Finally, we will merge this new data set with the suicide dataset indexed by years. It will have 3 columns: year, suicide rate, and proportion of negative headlines.

Below is an ideal dataset we aim to create from the samples (note: the numbers in the dataset are purely imagined, not the real results): *additionally the negative headlines % is calculated by our sentiment analysis that will be discussed in the next section

Year	Suicide Rate (%)	Negative headlines (Compound Score)
2011	15	0.0601
2012	12	0.0594
2013	23	0.0566

Text Analysis

To start the text analysis–specifically sentiment analysis, we need to start by organizing all the data in a pandas data frame so then It can be processed by our Natural Language Processing model (all of the preprocessing steps are outlined in the data wrangling section). To do all the processing for each of the New York times headlines, we are going to use an open source sentiment analysis that is called VADAR. This framework outputs a *Compound score* that is then associated with an emotional valence: positive sentiment: compound score >= 0.05, neutral sentiment: (compound score > -0.05) and (compound score < 0.05), negative sentiment: compound score <= -0.05. For our purpose we are going to use these scores to average the “sentiment” over a given time period with lower relative scores meaning a negative correlation and the inverse is True. After running this model for all of the New York Times headlines in the dataset each row of the data frame will be given its compound score that can then be used later for further analysis.*

*(Since we are using a premade NLP model, here is how it works) The process of natural language begins with Tokenization of the headlines, converting all the headlines to lowercase letters, removing all the stop words, and then removing all non words for the tokens. Then TF-IDF (Term Frequency-Inverse Document Frequency) is used to weight the words in the title based on the importance to emotion valence, and then Word embeddings and Bag of Words. Then go on to start training the model by supervised learning which then can identify our “Compound Score” of each of the Headlines.

Data Visualization

After we do all the data wrangling and the Text analysis we are going to create a line graph that has all the years on the x axis and then one line that is representing the Compound Scores and then a line that represents the suicide rate for that given year. We are going to adhere to the data to ink ratio principles.

Descriptive & Exploratory Data Analysis (summary stats, correlation, etc.)

With the suicide rate % and the compound score we are going to a simple regression to calculate the relationship between the variables and the associated R values to determine if the results can be used to paint a picture.

Discussion (10 pts)

Given your hypothetical results, how would you draw inferences / conclusion based off of those results? What are the limitations, pitfalls, and potential confounds of your methods, or biases in your data sources (e.g., how does the selection of the sources of your crowds affect your outcomes?)? How would you set out to address them? In addition, outline how you would address any societal and/or ethical implications of your proposed project discussed in your Ethical Considerations section.

Given your hypothetical results, how would you draw inferences / conclusion based off of those results?

If our results indicate a statistically significant correlation between the emotional valence of New York Times headlines and suicide rates in the U.S., we would infer that media tone may have a measurable impact on public mental health trends. A positive correlation would suggest that more negative headlines are associated with higher suicide rates, whereas a lack of correlation would indicate other factors are likely more significant drivers of suicide trends. Moreover, if a strong link was found, it would not mean that the tone of the headlines directly causes higher suicide rates. Instead, it would suggest that both negative headlines and suicide rates may be influenced by other factors like economic issues, mental health awareness, or social problems

What are the limitations, pitfalls, and potential confounds of your methods, or biases in your data sources (e.g., how does the selection of the sources of your crowds affect your outcomes?)?

One limitation of our methods is the use of the sentiment analysis model VADER has limitations in capturing complicated language, such as sarcasm or complex emotional undertones, which could affect the accuracy of sentiment scoring. Another limitation is that the New York Times has a specific audience and focuses on national and international issues, which may not represent the perspectives found in local or regional media. Headlines, being short, often do not capture the full tone or context of the articles, which can lead to incomplete or misleading sentiment analysis. VADER also relies on a fixed list of words, which can make it less accurate when interpreting specialized terms or unique phrases related to mental health reporting. These factors could lead to sentiment scores that don't fully reflect the true emotional tone of the headlines, affecting the reliability of the results.

How would you set out to address them?

To address these limitations, more advanced sentiment analysis tools, such as machine learning-based models like GPT-based frameworks, could be used to improve the understanding of complex emotions, sarcasm, and nuanced language that VADER struggles with. These models can be fine-tuned on datasets specifically related to mental health discourse, ensuring greater accuracy in interpreting specialized terms and unique phrases. To better capture the tone of the articles, we could look at not just the headlines but also the first few lines or summaries of the articles, which would give more context. Additionally, having experts review a sample of the headlines could help identify errors in the sentiment analysis and make the results more accurate.

In addition, outline how you would address any societal and/or ethical implications of your proposed project discussed in your Ethical Considerations section.

To address the societal and ethical implications of our project, we will prioritize transparency, accuracy, and sensitivity when examining the correlation between New York Times headlines and U.S. suicide rates. Due to the sensitive nature of suicide, we will ensure that our findings are presented with caution, clearly noting any correlations and making sure that they are clear. We will minimize bias by using a diverse and representative sample of headlines, avoiding political or sensationalized content that could distort our results. Additionally, by utilizing publicly available data and respecting privacy, we will adhere to ethical guidelines around data ownership and ensure that the information used is organized and anonymized. Recognizing the potential harm in misinterpreting the relationship between media sentiment and suicide, we will be careful to avoid drawing oversimplified conclusions.

Group Participation (3 pts)

Include one paragraph briefly outlining the contribution of each group member throughout the quarter while working on this project. Each of you must also fill out the survey (link provided toward the end of the quarter) about individual and group participation. **The results of this**

survey can negatively impact an individual's final grade if the group provides evidence that one member did not contribute to the project. (3 pts)

Throughout the quarter and our time spent together working on this project, it is safe to say that every group member has equally contributed towards the unified effort of this project. Starting from the beginning, Jaden was quick to spearhead the initial starting of our project by suggesting the overall topic for it. While all group members brainstormed together, Jaden was able to get the group started by proposing his idea. Moving forward, all group members were extremely eager and active in working on the project. While we did delegate various parts of the project to different group members, we always made sure to corroborate together as a whole and communicate with one another. Taking on various initiatives, Richard always kept everyone accountable and made sure that our level of communication was almost always at its highest efficiency. Leo was adamant on the organization and ethical methodologies of our project, never failing to take it upon himself to ensure this was maintained at all times. Yathin's research regarding background information was able to provide a solid foundation and framework for our project. And finally, as mentioned before, Jaden's initiative and overall strong efforts towards the project inspired everyone in the team, rallying us together to push forward. Working on this project together as a group has been almost inspiring even; seeing each member of the team take on a different lead role and coming together to put forth our best effort has been truly uplifting throughout our time together.