Final Project Progress Report

Abortion as a Public Issue: A Sentiment Analysis of Social Media versus Popular Media

1. Team Name: Team Zebra

2. Team Members (first and last name, email)

a. **Bhavya Pandey,** bhavyapandey@uchicago.edu
b. **Wenhao Li,** liwenhao@uchicago.edu
c. **Xiaochen Ding,** xcding@uchicago.edu
d. **Xiaowei Guo,** xiaoweiguo@uchicago.edu

3. GitHub Repository Link: https://github.com/orgs/macs30112-winter23/teams/team-zebra

4. Project Description

Objective of the Project and Rationale

The objective of our project is to study the difference in views about abortion, on social media versus in popular media between January 2022 and January 2023.
Social science research suggests that views about abortion can differ on social media and in popular media, with social media tending to amplify extreme views and popular media presenting a more nuanced view of the issue. It's important to be aware of these differences and to seek out a range of perspectives when engaging in discussions about abortion. Our project aims to help understand and deconstruct these differences through data visualizations, observations from data, and data modeling.

Social Science Perspective

Background, questions of interest to social scientists, recent developments, key concepts

Abortion rates have been declining in the United States. According to data from the Guttmacher Institute, the abortion rate in the United States declined by 7% between 2014 and 2017. Further, in terms of ideological assertions in public and political discourse in the US, there is a partisan divide on the issue of abortion. A Pew Research Center survey from 2019 found that 61% of Democrats believe that abortion should be legal in all or most cases, while only 27% of Republicans hold this view. Additionally, access to abortion is uneven across the United States. According to a report from the Guttmacher Institute, 27 states have policies that are hostile to abortion rights, such as mandatory waiting periods, mandatory counseling, and restrictions on insurance coverage. The decision to have an abortion is often a complex one. Research has shown that women who seek abortions do so for a variety of reasons, including financial instability, concerns about their ability to care for a child, and the desire to finish their education or advance their careers.

The question of whether abortion should be legal or not has led to the creation of two main groups: pro-choice and prolife. However, just recently, the Supreme Court of the United States (SCOTUS) officially overturned the Roe vs. Wade case on June 24, 2022. The Roe vs. Wade case first appeared in court in 1973 and was a landmark decision that stated that the Constitution of the United States granted citizens the right to an abortion. However, with the recent overturning, abortion rights are now decided by each individual state on whether it is a right or not.

Research Questions and Hypotheses

With the multitude of opinions in popular media and on social media, our project seeks answer to the following questions:

- How are people reacting to the topic of abortion online, particularly on Twitter?
- Are there any geospatial trends in where Tweets are posted from? Can we group states by their views on social media?
- What are the views being expressed through popular media, such as newspapers like the New York Times? Are views in popular media 'milder' (more neutral) than those on social media?
- Can the data from these observations be modeled or visualized to gather further insights and draw out parallels between the two channels of expression?
- Is there a difference between these observations after the June 2022 ruling?

Based on social science research in the past, we **hypothesize** that social media tends to amplify extreme views on abortion. On the other hand, popular media, such as newspapers like The New York Times, tend to present a more nuanced view of abortion. We expect a similar result after carrying out a 'sentiment analysis' on both sources (more details on this follow). While media outlets may have their own editorial biases, they often present a range of perspectives on the issue of abortion and the various political and social factors that contribute to it.

5. Data Sources

Data Source 1: Pregnancies, Births and Abortions in the United States: National and State Trends

To preface our analysis, we intend to provide a background into the historical geographical trend of pregnancies, births, and abortions in the United States.

- Type of data collection: downloading open source data available here
- Time frame: 1988 to 2017
- Size of the dataset: ~ 1000 observations across states over the years
- Possible data reliability/validity issues: none
- Additional information: this data is intended to support the rationale for our study

Data Source 2: Tweets related to abortion views

So far, we have collected a sample of data of around ~3000 Tweets, and will carry out preliminary analyses using this. In case we feel the need to further increase the robustness of our analyses, we will consider enlarging the sample size.

- Type of data collection: scraping using a Python library called `snscrape` that allows us to scrape Tweets by defining key words, time periods, geographical information and other advanced search parameters
- Time frame: January 2022 to January 2023
- Size of the dataset: ~ 3000 Tweets
- Possible data reliability/validity issues: since we are scraping the Tweet data by using textual, time, and geographic indicators, we need to be careful regarding how we define our scraping and searching conditions (for instance, we plan to collect all Tweets (replies and comments included) that include certain words such as 'pro-life' or 'pro-choice', and so on)
- Additional information: this data will then be used to conduct a sentiment analysis on the prevalence of views regarding abortion on Twitter.
- To ensure that the sentiment analysis is carried out accurately, since we hope the keyword to be tendentious, we intend to focus efforts on finding the right keywords for scraping. If we simply use 'abortion' as the keyword, it is hard to distinguish people's attitudes. For example, Tweets like "I hate people who are against abortion" are likely to be labeled as negative, but the actual attitude should be positive.
- One of the challenges we faced with this data source was finding a source that would allow us to get any number of geotagged Tweets. We tried a few options such as the Academic API, and other libraries in Python such as Tweepy, however these were not allowing the flexibility in data collection which we required. Finding the `snscrape` Python library has been helpful in this regard since its functionalities are quite advanced and it offers powerful and accurate search-based scraping of Tweets.

Data Source 3: New York Times articles related to abortion

In order to analyze the views on abortion in popular media, we intend to collect all articles published in the New York Times in the defined time frame and carry out the same sentiment analysis methodology used on the Tweets.

- Type of data collection: web scraping by integrating the NYT API
- Time frame: January 2022 to January 2023
- Size of the dataset: we are aiming for ensuring at least ~10 articles for analysis, per month, however the objective is to collect the text for every article published on 'abortion' in the specified time frame
- Possible data reliability/validity issues: we understand that this would be a very large dataset, so we are exploring whether we need to use the complete articles for our analyses or only the headlines/summaries
- Additional information: we are using the keyword 'abortion' to scrape the relevant articles.

- In order to ensure that the journalistic bias of the NYT does not skew the results, we may consider scraping another news source if we feel that it would benefit the objectivity of our project.

6. Data cleaning/wrangling

- Since we are dealing with a lot of textual data, we are using Pandas dataframes as our basic data structures for storing and processing the data.
- Most of the data cleaning would be done using the functionalities of Pandas, and tools such as regular expressions integrated with the libraries used for data collection.
- \`snscrape\`, the main library we are using for Tweet scraping takes care of many of our concerns regarding data cleaning of the Tweets collected.
- We have written functions to clean and pre-process our Tweets for analysis.
- Using methods as discussed during the course, we intend to pre-process, format, clean, and store the textual data collected from the New York Times in a manner that allows us to run the sentiment analysis on it.

7. Data analysis and visualization

Details Regarding the Sentiment Analysis

- We plan to conduct sentiment analysis using Data Sources 2 and 3, as described above.
- We are using a pre-trained model for this analysis, called the RoBERTa-Base Model. This transformer-based model was trained on roughly 124 million Tweets from January 2018 to December 2021 and is specifically fine tuned for sentiment analysis involving media sources (like social media and newspapers) which makes it ideal for our project.
- This model comes with its own classification labels and tokens, which classify the Tweets/text into Positive, Negative and Neutral sentiments.

Analysis and Visualization Elements

- As a preface to the topic tackled through our project, we will first present a historical state-wise analysis of births, pregnancies, and abortions over the years in the United States (reference: Data Source 1)
- Firstly, we want to build an animated map that shows the attitudes to abortion in the last one year, in each US state. We collect tweets with geographic labels through the method described above. Then perform sentiment analysis on Tweets in different states/regions, we will use a pre-trained model, and the Tweets will be divided into positive, negative, and neutral categories.
- To build the map, we need to calculate (the number of positive sentiment tweets/the number of negative sentiment tweets) for each state. In this way, we can get lists of each state's attitude toward ‘anti-abortion’ for the time frames described, and store these as a score.

- These scores can then be utilized for constructing the interactive map visualization, as described by our group member Wenhao Li in the class lectures pertaining to HW4.
- With regards to the NYT sentiment analysis data, we plan on calculating the score for each article (the number of positive sentiment articles/the number of negative sentiment articles) in the specified time frame and then draw a trend graph of the media's views on abortion according to the proportion of positive sentiment, in a monthly manner.
- Furthermore, we also intend to normalize the two scores over the time frame, in a monthly fashion, and study the progression of sentiment in social media versus popular media over this period. In particular, we intend to discern if there difference between these observations after the June 2022 ruling.
- We also intend to use these normalized scores for further exploratory data analysis (including descriptive statistics and simple visualizations), and even use some modeling techniques to further deconstruct the 'causality' of the observations we have drawn so far. Some options we are currently considering are:
    - Linear regression between the social media and popular media scores (considering both cases of either being dependent on the other)
    - Time series analysis of the two types of scores

<u>Python Libraries for Data Management, Visualization, and Modeling</u>

- NumPy and Pandas for data management
- Transformers, PyTorch, and SciPy to implement the sentiment analysis model, import its labels, and run it on the data
- Time for time formatting
- Statsmodels for data modeling and statistical analyses
- Requests and BeautifulSoup for web scraping
- Matplotlib and Seaborn for basic visualizations
- Plotly.express for mapping

8. Responsibilities

1. **Bhavya Pandey:** Twitter scraping and implementing sentiment analysis, documentation
2. **Wenhao Li:** Mapping, statistical analysis and interpretation
3. **Xiaochen Ding:** NYT scraping and implementation of sentiment analysis
4. **Xiaowei Guo:** statistical analysis, visualization, data modeling and interpretation

9. Selected References

- Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. ACM Computing Surveys (CSUR), 49(2), 1-41.
- Han, B., Cook, P., & Baldwin, T. (2014). Text-based twitter user geolocation prediction. Journal of Artificial Intelligence Research, 49, 451-500.

- Unified Benchmark and Comparative Evaluation for Tweet Classification, Barbieri, Francesco and Camacho-Collados, Jose and Espinosa Anke, Luis and Neves, Leonardo. Findings of the Association for Computational Linguistics: EMNLP 2020. 10.18653/v1/2020.findings-emnlp.148, 1644--1650.