Final Project 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/macs30112-winter23/final-project-team-zebra

4. Project Description

Project Goal: 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 and Brief Literature Review

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 **hypothesized** 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 [more details about methodologies, validity checks, pre-processing in the Jupyter Notebooks and slides]

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

As a preface to our data collection and analysis using Twitter data and NYT data, we wished to provide the context regarding the historical trend of Pregnancies, Births, and Abortions in the United States from 1988 to 2017. The objective of visualizing this data obtained from a publicly available dataset by the Guttmacher Institute, is to establish the way that abortion has been viewed

as a public issue across geographical regions, over time, and how the incidence of births, pregnancies, and abortions in different places at different times, reflect this notion.

- Type of data collection: downloading open source data by the Guttmacher Institute 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

Initially, we collected a trial sample of data of around ~3000 Tweets, and carried out preliminary analyses using this. Eventually, we have increased the number of Tweets to ~8500 to make the analysis more statistically significant.

- 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: ~ 8500 Tweets
- Tackling data reliability/validity issues: During the data collection process, we tried a few combinations of keywords to scrape Tweets as well as NYT data, and after a few trial-and-error instances, we settled on the methods which we deemed were best for the analysis. Since we scraped the Tweet data by using textual, time, and geographic indicators, we were extremely careful regarding how we define our scraping and searching conditions (for instance, we plan to collect all Tweets (replies and comments included) that include the keyword 'abortion', between our defined time period, and was from a particular geographical location).
    - We ensured **internal validity** in our data collection process by ensuring a singular, common keyword for scraping
    - The time periods were established, and data collected individually so that no inherent biases creeped into the data collection process
    - Further, the coordinates of location and a defined radius around them were used to maintain consistency in terms of the geocodes being used for scraping the Tweets
    - We have also used methodology and code in such a manner that the **external validity** of the collected data is also ensured
- 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 focused efforts on finding the right keywords for scraping. This effort was ratified by the deployment of the RoBERTa-Base Model for sentiment analysis which is

a very sophisticated pre-trained model that reads the content in between the lines to ensure the correct labeling and classification of the text.

- 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 that pertain to the word 'abortion' and carry out the same sentiment analysis methodology used on the Tweets, on the leadbody and titles of these articles.

- Type of data collection: web scraping by integrating the NYT API
- Time frame: January 2022 to January 2023
- Size of the dataset: ~500 articles
- Tackling data reliability/validity issues: since the official NYT API was used for scraping, we were not very concerned about the reliability of the data and its validity for this data source. However, since this was a large amount of textual data, we made sure to pre-process and clean the data appropriately in order to make sure that the final sentiment analysis using the RoBERTa-Base Model be carried out accurately on it.
- Additional information: we are using the keyword 'abortion' to scrape the relevant articles.

6. Data cleaning/wrangling [incorporated in the Jupyter Notebooks]

- 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.
- TextBlob and other libraries have been used to clean and pre-process the large textual data as well
- `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 and article blurbs for analysis.

7. Data analysis and visualization [incorporated in the Jupyter Notebooks]

Details Regarding the Sentiment Analysis

- We have carried out sentiment analysis using Data Sources 2 and 3, as described above.

    - The first method we used for this was using the NLTK toolkit library in Python, however this had its limitations in terms of not appropriately classifying some of the nuanced text which was a part of our datasets (more details on this are available in the Jupyter Notebooks).
    - Hence, we carried out the final sentiment analysis using the method described below.
- 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)
- With regards to the sentiment analysis data from sources 2 and 3, it has been analyzed and visualized in the following ways:
    - Using a Bubble Map to show the geographic dispersion of the Tweets's sentiments
    - Using word clouds
    - Plotting of time-wise trends
    - Plotting of distributions
    - Plotting of descriptive statistics
        - And more…
- These analyses and visualizations can be found in the Jupyter Notebooks in the dedicated folders for the two main data sources, as well as, the dedicated folder for the data visualizations.

Main Python Libraries Used for Data Collection, Management, Visualization, and Modeling

- Snscrape for scraping the social media data
- Sklearn, NLTK, and TextBlob for running machine learning based sentiment analysis methods on textual data (NYT data in our case)
- NumPy and Pandas for data management
- Transformers, PyTorch, and SciPy to implement the pre-trained sentiment analysis model, import its labels, and run it on the data
- Time, datetime, dateutil for time formatting
- Statsmodels for data modeling and statistical analyses
- Requests and BeautifulSoup for web scraping
- Matplotlib and Seaborn for basic descriptive visualizations
- Plotly.express for mapping
- WordCloud for word clouds
- Collections for keeping track of information in code

8. Key Findings, Takeaways, and Future Scope of Work

Results of the sentiment analysis showed that:

- Overall, Tweets show a greater extent of extreme sentiments – particularly negative sentiment, across all four major cities we studied
- On the other hand, the New York Times articles (both headlines and lead bodies) generally showcase 'milder' sentiments, such that a more neutral sentiment can be observed for them, across the time-period we studied (important to note here that positive sentiment is not observed in both cases)

Effect of the June 2022 ruling:

- Our analysis showcase that the effect of the 2022 ruling was evident in the way the word 'abortion' appeared in both, Tweets and NYT articles: while there was an increase in volume of both, Tweets as well as articles, the net effect was convergent to the effect seen in general: negative sentiments for Tweets, and neutral for articles
- There can be observed a peak in negative sentiment Tweets at the time of the overruling, and a synchronous peak of neutral sentiment headlines in NYT around this same time period - which ratifies our conclusions so far

Geographic incidence of views on social media:

- We also observe that there is a difference in the way sentiments are expressed, on a geographic basis. We studied 4 large US cities. The number of positive sentiment Tweets were distributed as: New York > Chicago > Seattle > Los Angeles

For this component of the project, we also attempted to undertake certain quantitative modeling techniques to yield more insights regarding the data, however since the sentiment analysis scores computed by BERT models are probabilistic in nature, it was not feasible to derive results from these and include them as a part of the project. We believe that data visualizations and trend analyses were appropriate and apt ways of analyzing the data we collected, and hence these eventually formed key components of our project.

We also had some great takeaways from the project as a team, and as budding computational social scientists! Some of these include:

- ★ Using social media data effectively – scraping using various channels, pre-processing, and examining multiple use cases to ensure that the data quality is maintained + no biases creep into the data collection process
- ★ Using data from web APIs and scraping – cleaning it, storing it properly for efficient usage, and extracting best use cases from it
- ★ Looking for loopholes and possible shortcoming at each step of the data management process, course-correcting along the way
- ★ Experimenting with different visualization techniques, storytelling using data, gathering insights

★ Other important experiences and learnings, that we believe are worth mentioning:
    - Coding as an effective team
    - Iterating over problems
    - Finding solutions together
    - Sharing common learnings and experiences within the team

There is still great scope to expand the work we have done, such as:

- Causal inference from geographic incidence of views is an aspect of our research question which can be explored further in the future
- Expanding the sentiment analysis to longer time periods, regions, social media platforms, and popular media platforms
- Juxtaposing the official statistics of births, pregnancies, and abortions with the corresponding sentiment analysis of social and popular media to further study the relationship regarding how real-world decisions of people are reflected through media channels, if at all
- A differences-in-differences regression discontinuity design using the sentiment scores can be explored further to understand the causal effect of the June 2022 overturning
- A longer look at the data can allow us to come to a more robust understanding of the trend which can be further explored through more longitudinal visualizations by looking for further data from relevant time periods

Furthermore, some limitations of our work include:

- Biases can creep into our findings since we have collected data from urban centers only.
- The BERT model we have used is computationally expensive, and since it comes with predefined label criteria, it can sometimes be hard to interpret its results accurately.

9. Responsibilities

1. **Bhavya Pandey:** Twitter scraping and implementing sentiment analysis, literature review, documentation, interpretation and presentation of the analysis
2. **Wenhao Li:** Data visualization and interpretation, documentation
3. **Xiaochen Ding:** NYT scraping and implementation of sentiment analysis, data visualization
4. **Xiaowei Guo:** Twitter scraping and implementing sentiment analysis, data visualization and interpretation

10. 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.
- NLP Modelling using RoBERTa-Base Model, video tutorial available here.
- Pregnancies, Births and Abortions in the United States: National and State Trends by Age dataset available here, and featured in a report by the Guttmacher Institute here.
- Pew Research Center's survey on abortion, available here.