A 30-Day Sentiment Analysis on the keyword 'Abortion' from Twitter USA

In partial fulfillment of the requirements for
CSE 482 - Big Data Analytics
taught by Dr. J.T.

by
Gacis, Angelica Louise M.
Hong, Yena

Masters of Science in Data Science
College of Natural Science
Michigan State University

December 16, 2022

1. *Introduction*

Earlier this year, the ruling for Roe v. Wade was overturned. This means that the legislation making abortion a federal right in the United States is now ineffective. An uproar of an already existing debate regarding abortion ensued as a result of several states immediately banning abortion soon after the ruling was implemented (Housman, 2022).

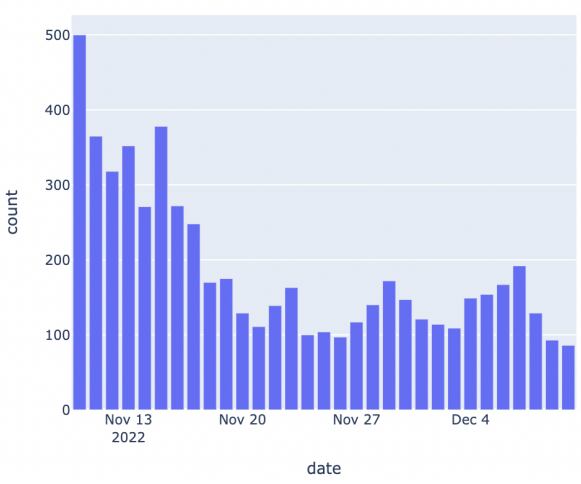
The aim of this study is to quantify the sentiments of people in the United States through their Twitter tweets regarding the topic of abortion. This will be done by the construction of different classifiers wherein the classifier producing the highest accuracy will be chosen for the final prediction and analysis of results.

2. *Solution*

2.1 Data Collection

We used the free 30-Day Twitter API Search or Sandbox subscription for scraping the tweets through the `tweepy` package. The keyword used is "abortion" and retweets were not included. The data collected is filtered only to tweets between November 10 to December 10 in the year 2022 and geotagged in the US. The summary number of tweets are shown in the plot below.

Figure 1. Scraped Tweets per Day



2.2 Data Cleaning

This study seeked the help of the `nltk` package for cleaning the scraped tweets. First, the '@' sign was removed from all the tweets with mentions, for example. Other punctuations, numbers, and special characters were also removed. We also removed stop words since they are not of much help in identifying a text's sentiment. All tweets were then converted to lowercase and tokenized. Lastly, stemming, specifically the stripping of suffixes, was applied to all tweets.

The variable indicating the place of where each tweet was tweeted was a list of dictionaries. Only the state and the bounding box coordinates were extracted from the dictionaries. From the coordinates of the bounding box, we calculated the coordinates of the center.

2.3 Initial Tagging

We created a "sentiment" variable and manually assigned "positive", "negative" and "neutral" to some of the tweets for the training of the machine learning models. We also used the help of the nltk package called SentimentIntensityAnalyzer to improve our work for the tagging of 70% of the data.

2.4 Supervised Machine Learning

2.4.1. Feature Engineering

We applied the bag-of-words method from `scikit-learn` to transform our string data to count vectors which the machine learning models can easily read.

2.4.2. Model Selection

The features generated from the bag-of-words algorithm were used in four different estimators namely, Logistic Regression Classifier, Random Forest, SVM Linear, and Naive Bayes with accuracies of 94%, 69%, 77%, and 76% respectively. The best model, the Logistic Regression classifier is then used to predict the sentiments in the remaining 30% of the data.

2.5 Data Visualization

There are three main visualization tools used to present the results of the data. First is the time-series plot via `plotly`. One time-series plot shows the counts of tweets with positive, negative, and neutral sentiments per day. Another time-series plot shows us the net sentiment per day. This was computed by subtracting the counts of negative tweets from the counts of the positive tweets.

The `pydeck` package helped us create a map of the USA with a scatterplot showing the points where each tweet was located. Green dots are for positive sentiments regarding abortion, yellow dots are for neutral sentiments, and red dots are for tweets indicating negative sentiments regarding abortion.

The last visualization tool is the word cloud which summarizes the strongest words mentioned regarding abortion. We included one word cloud for the positive and another for the negative tweets. This was done with the `wordcloud` package.

With `streamlit`, we have also created a web app that features all of these visualizations where the user can also filter the data based on a specific date or state in the US.

Web App:
https://almgcs-bigdataanalytics-cse482-uqbv8c.streamlit.app/

3. *Results and Discussion*

Using the Logistic Regression Classifier as our final estimator, we were able to create a new column we called “y_predict” which tags each cleaned tweet as either “positive”, “negative”, or “neutral”.

3.1 Most Frequent Terms

The green wordcloud below shows us the most mentioned words from cleaned tweets whose sentiments are predicted to have positive sentiments on abortion by the Logistic Regression. The size of the word indicates how often it is mentioned. Besides the word "abort", the most mentioned words are "people", "want", "right, "women", "state", "vote", and "support". There are also noticeable words or phrases such as "republican", "democrat", “georgia”, "care", "pass", "know", "live free", "protect", "allow", "decision", "help", "freedom", and "love". We can see that most of the tweets with positive sentiments on abortion talk about the right to decide and live free as well as the protection of women.

Figure 2. Positive Sentiment WordCloud

Words like "women", "right", "state", "vote", and "people" also dominate the red wordcloud or those clean tweets predicted to have negative sentiments on abortion by our most accurate classifier. However, it is noticeable that the word "kill" is mentioned a lot in these tweets. Words that also came up especially for this wordcloud are "murder", "trump", "wrong", and "ban". Political terms such as "republican" and "democrat" are mentioned less in anti-abortion tweets than in pro-abortion tweets from our 30-day data.

Figure 3. Negative Sentiment WordCloud

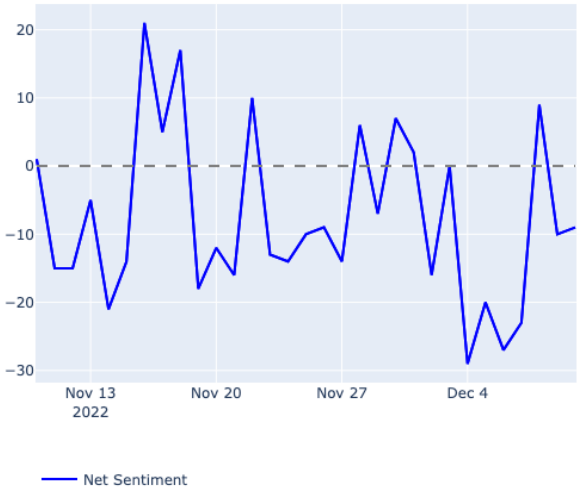
3.2 Sentiments over 30-day period

We can see from Figure 5 below that the total number of original tweets in the US mentioning the word “abortion” started out abundant until it slowly decreased towards the end of November and started to increase again by the start of December. The abundance of

tweets at the start may be due to lingering Twitter discussions on abortion following the recent elections on November 8 this year whereas our data collected started on November 10.

Figure 5. Total Time Series per Sentiment

Figure 6. Total Time Series for Net Sentiment



The figure 7 and 8 below indicate the total time series for positive and negative sentiments grouped by five regions of the United States: Midwest, Northeast, Southeast, Southwest, and West. From the two sentiment graphs, we can notice that there is no distinct difference among the portion of each region per day. In other words, the patterns of each region may follow the pattern of the whole United States.

Figure 7. Total Time Series for Positive Sentiment by Region

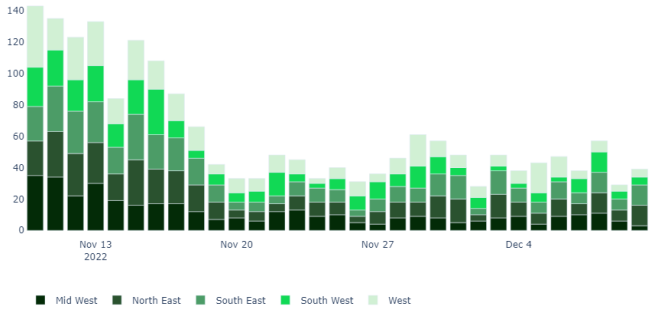
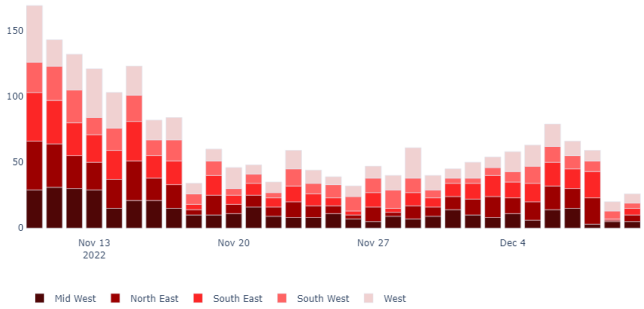


Figure 8. Total Time Series for Negative Sentiment by Region

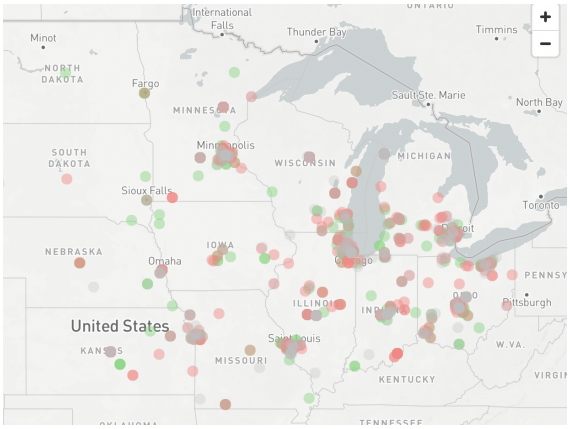


3.3 Sentiments per Area

The figures from 9 to 13 depict the overall sentiment of the five regions by locating longitudes and latitudes of tweets on the map with color green, red, and gray. Each color represents positive sentiment, negative sentiment, and neutral sentiment respectively. Although, the number of positive sentiments and negative sentiments approximately equal, we can find the regional characteristics from the figures.

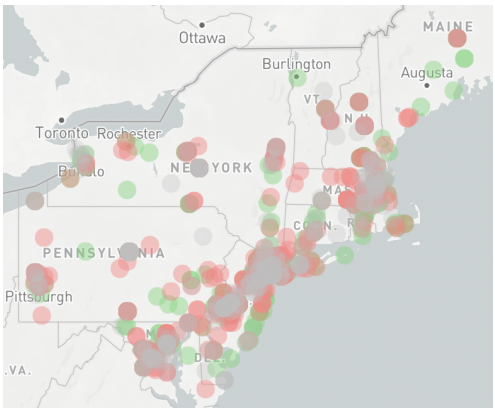
In the Midwest region, we can see positive sentiments relatively frequent on the west side, and the negative sentiments on the east side.

Figure 9. Midwest Region Sentiment Map



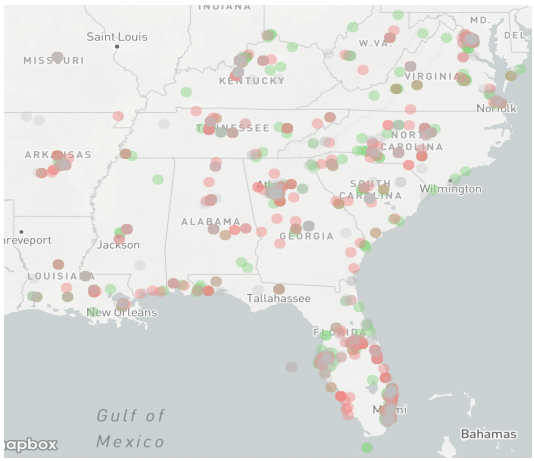
In the Northeast region, most of the positive, negative, and neutral sentiments appear along the coast line.

Figure 10. Northeast Region Sentiment Map



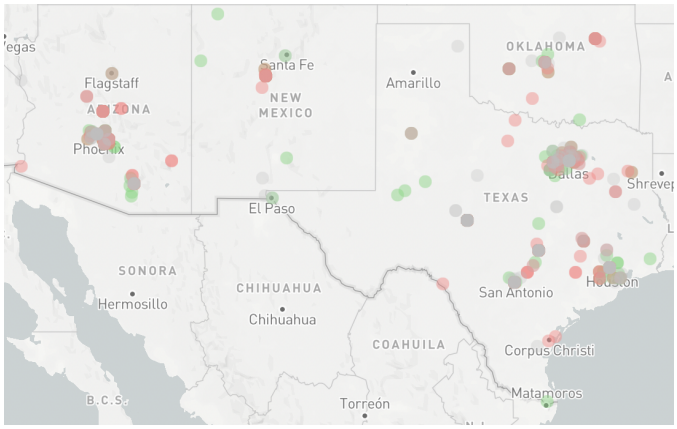
In the Southeast region, the negative and neutral sentiments are relatively heavily clustered in Florida, but are generally spread throughout the Southeast region.

Figure 11. Southeast Region Sentiment Map



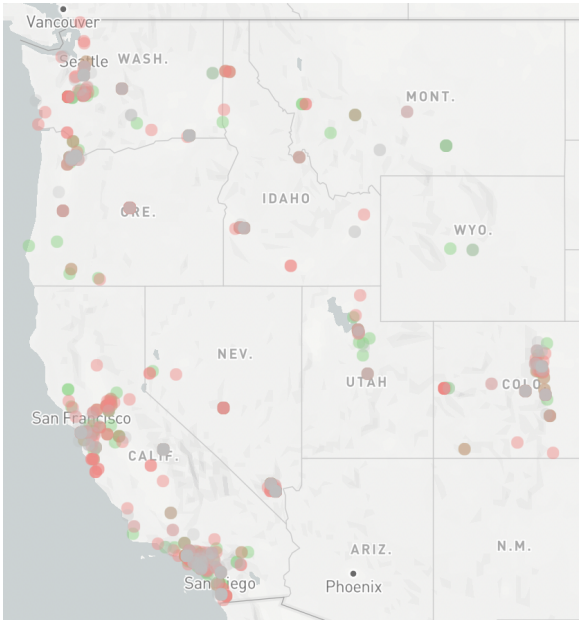
In the Southwest region, most tweets were from Huston and Dallas in Texas and Phoenix in Arizona. The negative sentiments appear more often than the positive sentiments.

Figure 12. Southwest Region Sentiment Map



In the West region, most tweets originated from California, and there seem to be more neutral and negative sentiments than the positive sentiments.

Figure 13. West Region Sentiment Map



Generally, the number of tweets per area are still proportional to the population in those areas. We can see several positive, neutral, and negative tweets per region which indicates that the sentiment of people per area on abortion may not be as unified yet regardless of whether it is legalized by their state government or not. With this, we can infer that there may be other factors

which can affect their sentiments such as perhaps, religion or age.

Figure 14. Summary of Sentiments per US Region

Region	y_predict	
MW	negative	426
	neutral	278
	positive	369
NE	negative	431
	neutral	295
	positive	411
SE	negative	491
	neutral	345
	positive	462
SW	negative	363
	neutral	217
	positive	329
W	negative	480
	neutral	352
	positive	383

4. *Conclusion, Lessons, and Recommendations*

Tweets including the keyword 'Abortion' during the past month show that Twitter users from the US have posted their opinion actively when the statewide general election was held. As the time went by from the election date, November 8, the users were less likely to tweet about 'Abortion'. Since our dataset starts three days after the election, we could not analyze Twitter sentiments about abortion before the election. Therefore, we may extend the length of the observation period in our further study given more time or using a different type of subscription.

The supervised learning methods we utilized in the project may not be the optimal method. To be specific, classification algorithms that we could apply are not limited to Logistic Regression Classifier, Random Forest, SVM Linear, and Naive Bayes. We can try other different algorithms in our further study.

In addition, machine learning classifiers relied on the initial tagging that we made. However, we should also note that our perception of the tweets' sentiments may not reflect the actual sentiment of the original user. We also noticed that the nltk tool called SentimentIntensityAnalyzer that helped us with the manual tagging of 70% or around 4,000 of the cleaned tweets seemed to have difficulty in recognizing sarcasm. Several tweets from the dataset included sarcastic questions or phrases and some tweets which have positive sentiments towards abortion, but are stated in seemingly aggressive tones, were also tagged as negative. Hence, we would suggest taking more time in the manual tagging for the model improvement.

From the total time series for net sentiment, most of the net points of each day are below the baseline equals to zero. We can infer that the more users tend to tweet about negative sentiment rather than positive sentiment on abortion. In the further study, we can focus on finding out whether the negative sentiment tweets were resulted from the unique number of users or the same users.

5. *References*

Housman, P. (2022) *Roe v Wade overturned: What it means, what's next*, *American University*. Available at: https://www.american.edu/cas/news/roe-v-wade-overturned-what-it-means-whats-next.cfm

Hunt, Lauren & Goldstein, Carly & Garnsey, Camille & Bogen, Katherine & Orchowski, Lindsay. (2022). *Examining Public Sentiment Surrounding Abortion: A Qualitative Analysis of #YouKnowMe.* Women's Reproductive Health. 9. 10.1080/23293691.2021.2016161.

Mane, Heran, Xiaohe Yue, Weijun Yu, Amara Channell Doig, Hanxue Wei, Nataly Delcid, Afia-Grace Harris, Thu T. Nguyen, and Quynh C. Nguyen. (2022). *Examination of the Public's Reaction on Twitter to the Over-Turning of Roe v Wade and Abortion Bans.* Healthcare 10, no. 12: 2390. https://doi.org/10.3390/healthcare10122390

Samal, Biswaranjan & Panda, Mrutyunjaya & Behera, Anil. (2017). *Performance Analysis of Supervised Machine Learning Techniques for Sentiment Analysis.* 10.1109/SSPS.2017.8071579.