Data Science - Fall 2022 - Mini Project III

Aicha Moussaid
Aicha.moussaid@abo.fi
EDISS Master's Programme - Åbo Akademi University, Turku, Finland

Introduction

Abortion has been addressed from a machine learning perspective as a healthcare and social issue to study. Researchers have either attempted prediction of recurrent spontaneous abortions [1], systematic reviews on improving pregnancy outcomes [2], or predicting attitudes towards abortion [3]. With this significant existing background, this study's focus sheds light on how people and organizations talk about abortion on social media, specifically Twitter, by examining tweets related to the hashtag #**RoeVWade**.

Before diving deeper, an explanatory context comes in handy to put in value the nature of this work and the significance of its outcome. On January 22, 1973, the Supreme Court issued a 7–2 decision in favor of "Jane Roe" (Norma McCorvey) holding that women in the United States had a fundamental right to choose whether to have abortions without excessive government restriction and striking down Texas's abortion ban as unconstitutional [4]. This constituted Roe v. Wade. Although, on June 24, 2022, the Supreme Court ruled in Dobbs v. Jackson Women's Health Organization (JWHO). The ruling upheld Mississippi's ban on abortion at 15 weeks of pregnancy, overturned Roe v. Wade, and ended the federal constitutional right to abortion in the United States. This overturning erased 50 years of precedent, and triggered a strong public reaction in the U.S.

This project is being conducted in order to analyze these reactions, by outsourcing it to Twitter data in order to gain a general perspective. The topic received strong responses out of Twitter users which made the hashtag #**RoeVWade** trending. Studying the overall interaction with such a matter through this hashtag, from a data scientist's perspective, could report interesting insights. The analysis of the hashtag is done on Tweets data by setting target questions with potential discussion pools, and then network analysis along with visualizations and sentiment analysis are performed to answer them accordingly. This study is not meant to communicate any personal judgement of the situation but would instead observe the reaction of a neutral unbiased machine.

Data Collection

The data collection of this project was realized through the use of Tweepy. Tweepy is an open-source Python package that provides a very convenient way to access the Twitter API with Python. First, a Twitter Developer Account was setup, then an app is created for the project which will give the access to the keys that allow access to data using the Twitter API. These keys consist of: *API key*, *API key secret*, *API access key*, and *API access secret*.

The next step after is to use Tweepy to access the Twitter data. The implementation

aimed to get five thousand tweets since the date of the overturning of Roe v. Wade, June 24, 2022. The target information scraped were the ***username***, the ***creation date*** of the account, the ***verified or not*** status of the account, the ***description*** on the user's account, the ***location*** of the user, their ***following count***, their ***followers count***, the ***total tweets*** of the account, the ***retweet count*** of the tweet, the ***text*** of the tweet, and the ***hashtags*** mentioned in the tweet. The totality of the tweets ended up amounting to 4519 tweets, the language was preset to English, and they were saved in "**scrapped_tweets.csv**".

After the collection of data comes the cleaning, which is done through searching for duplicates in the rows based on the usernames, to not get retweet posts of previous tweets existing in the database already. This narrows down our dataset to 3764 unique tweets. The Unnamed column was also dropped since there will be no need for the enumeration of the rows at hand and will cloud further visualization.

The data cleaning will later on be performed in the sentiment analysis section to clean the texts of the tweets, since they can contain mentions, hashtags, links, punctuations, and many other things. And when working on a machine learning or data science project, this is a crucial step before processing them any further.

Data Analysis

This section is dedicated to analyzing the data previously collected and using it accordingly to answer specific questions. The questions are formulated into three main ones:

1- **Where is this hashtag mostly located?**
2- **When were the user accounts tweeting created at?**
3- **What were the highest interactions with the tweets posted and what sentiment did they generate?**

Answering these questions can help us investigate and gain perspective on three main sides of this analysis:

- Knowledge about whether the Roe V. Wade overturn triggered worldwide interaction or just a local one.
- Knowledge on whether there were fake accounts/bots that could influence and control the overall public response, which has been done previously to falsify the voice and opinion of the citizens.
- Get insight on who is behind the tweets (Verified account or not) with most interactions and what sentiment do the tweets hold.

RQ1: Where is this hashtag mostly located?

To answer this question, network analysis was adopted to clearly visualize the main locations the tweets are coming from. NetworkX was used implementation-wise for its ability to create, manipulate, and study the structure, dynamics, and functions of complex networks. It is worth noting that the initial dataset collected contained different values for the ***location*** field. In other words, since the location information is something the account user sets for himself, and could contain anything (emojis, flags, characters of other languages...), it was hard to find a uniform way of creating

set locations without getting involved manually.

By examining the data, and visiting the users main Twitter page, finding main country names was made a lot easier. We could clearly find ones that belonged to the US since the majority had the state name, or the state name accompanied with USA in the end (Florida or Florida, USA), other country names were also clearly stated like France, Canada, Australia, UK… There were also cases where the location will have multiple "?????". These were cases of emojis of flags, or characters in Japanese or Korean, that state the name of the country clearly. For the rest of the cases, a lot of the users had blank or random expressions set as their location like "Earth", "Rent free in your mind", and some even could notice what the study was doing and clearly stated "Stop data mining!". These cases were set to be "Unspecified". Excel provided a lot of the help here with its built-in functions where you could sort the values alphabetically and can of course duplicate the value on multiple rows without having to write "USA" hundreds of times.

By the end of this tedious part, our new dataset "**tweets.csv**" was ready to be handled and fed to a network. In total there were forty-four location nodes and 3980 edges with username nodes. The location labels consisted of: USA, Unspecified, Canada, UK, Australia, France, India, Germany, Netherlands, Peru, Puerto Rico, South Africa, Colombia, Israel, Italy, Japan, New Zealand, Philippines, Singapore, Sweden, Switzerland, Venezuela, Argentina, Bahrain, Brazil, China, Denmark, Ecuador, Finland, Iceland, Kenya, Mali, Mexico, Morocco, Poland, Portugal, Romania, Russia, South Korea, Spain, Thailand, UAE, Ukraine, and Vietnam.

This resulted in a graph of 4024 nodes and 3980 edges. We were able to get top nodes by degree, which had USA, Unspecified, Canada, and the UK at the very top, followed by the rest of the countries. The same results were displayed when ranking them by betweenness centrality, which gave the countries with only one edge (one username from that location) a value of 0.0. Communities were also made and for each class, it was shown that the name of the countries had the highest Eigenvector Centrality. There were only twenty-two classes though, stopping at Venezuela, since the rest of the countries had only one edge to them, and did not differ from the regular nodes of usernames, which have no particular transitive influence on them as nodes.

The visualization part was very messy and unreadable from NetworkX, and for better understanding and clear illustrations of the network created, Gephi was used. We imported "**tweets_network.gexf**" from Google Colab to Gephi. The nodes were partitioned and ranked by degree and Yifan Hu Proportional was used as a layout, which helped us obtain the following:

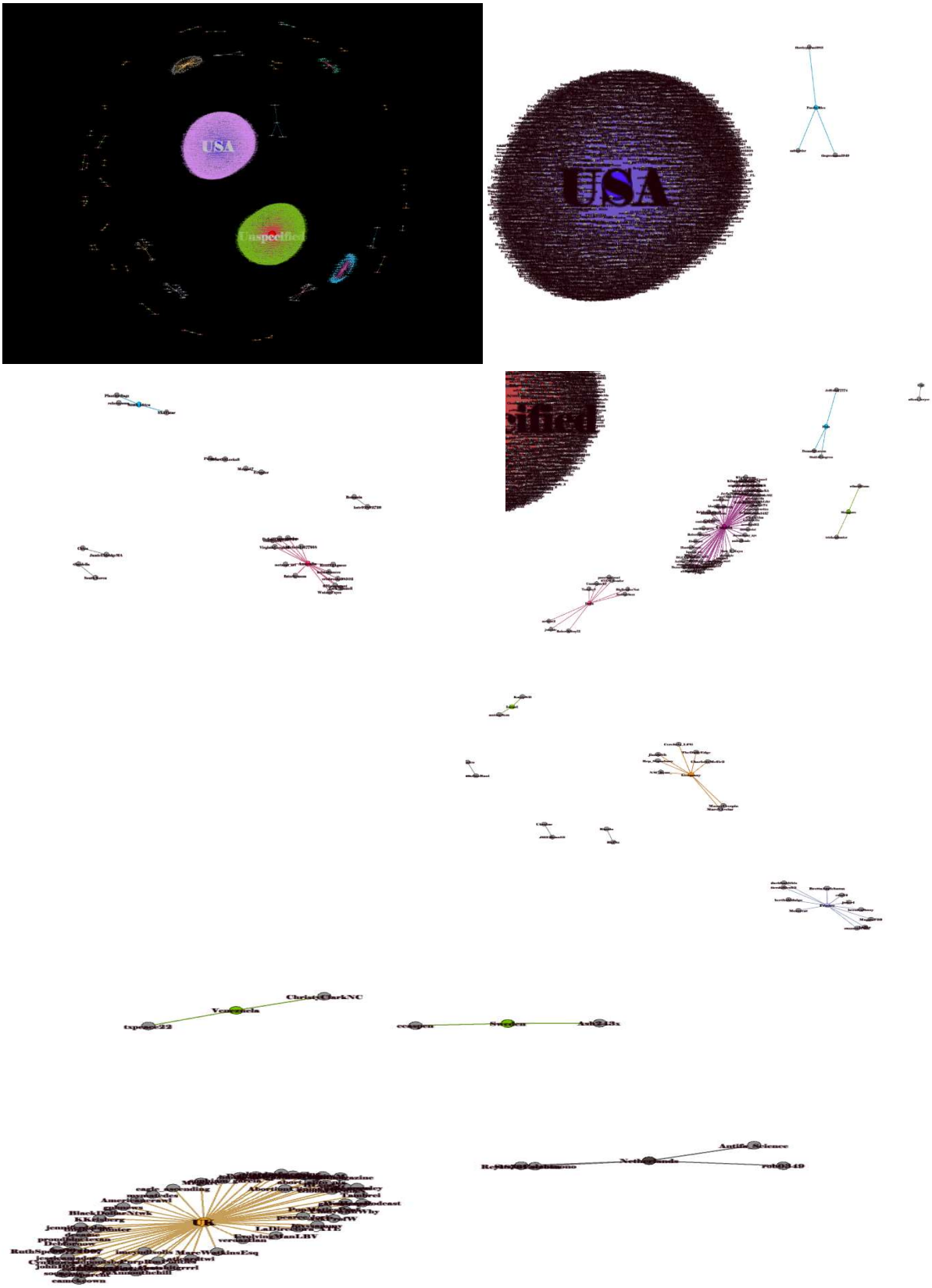

Figure 1. Snapshots of the location network graph from multiple focus points

From Figure1., we can clearly visualize the graph in a more understandable way where we get to see the clear clusters of nodes and the degree partitioning of each location. We can see how the entirety of the network has a galaxy outline, where the two main nodes are USA and Unspecified, as they have lots of edges linked to them from multiple Twitter users (2119 edges). Next comes Unspecified with 1642 edges, making it the second most dominant location node. We can also see the UK, Canada, Australia having their own smaller yet significant clusters, and we move to much lower nodes degree-wise, like Sweden, and Venezuela, and one-edged nodes like Morocco, Japan, and South Korea.

The Gephi plot helped to clarify the overall view and contents of our dataset and gave us a clear idea on how to answer the first research question of this study.

#RoeVWade was scattered around forty-four locations, and since most of the Twitter users also fall under the Unspecified category location-wise, accounting for 41% of black-boxed data, there is no straightforward way of knowing whether the number of these locations extends more. Although, with the data scrapped and the sufficient info gathered from the Network, it is safe to assume that the main concerned parties with this hashtag, and the ones reacting the most were U.S citizens. It is only logical for the people of the country to react to a law change in their own legal system, limiting one of the main liberties in the most liberal country in the world. This does not discredit the 5% reactions scattered around the globe, but the percentage of the global reaction still is minor compared to the local one. It is possible to speculate on the non-local responses and either explain it by assuming those accounts and tweets belong to U.S expats, or that the tweets are public worldwide opinions from different people interested in the cause behind the law, the repercussions of overturning Roe v. Wade, and expressing their views on global news in general.

RQ2: When were the user accounts tweeting created at?

This research question was the easiest to answer. It was inspired from the **#bolivianohaygolpe** hashtag study case, where it was found that lots of bots were generated to publish tweets that hogged the citizens genuine responses and opinions [5]. Its implementation did not take much effort since it was easily answered with a simple query of duplicate values based on the "Creation Date" of the user accounts. It returned no values at all, which led us to manually checking through Excel. Using the filtering feature by column in Excel allowed us to sort the data by values and therefore check for any obvious duplicates in that column. This technique is useful for identifying trends such as account name generators and same account creation dates. In the case of this hashtag, there were no suspicious accounts that could lead us to check whether they are bots or not.

Even though the dataset is relatively small to judge the entirety of Twitter user's reaction, we can assume through this analysis that the public reaction was genuine and did not involve any outsider intervention that could control the public opinion or falsify the claims online about the overturning and its repercussions.

RQ3: What were the highest interactions with the tweets posted and what sentiment did they generate?

The goal of this question was to get a general idea on the contents of the tweets collected and getting to know what the general, widely-adopted opinion in those tweets is. Firstly, the analysis started with generating a heatmap (Figure 2.) of "scrapped_tweets.csv" cleaned data, which illustrates the different correlations between the columns. From Figure 2., it is shown that there is a correlation between the number of followers and the Verified status of the account. Although, it does not portray any signs of that status affecting the retweet count of the posts.

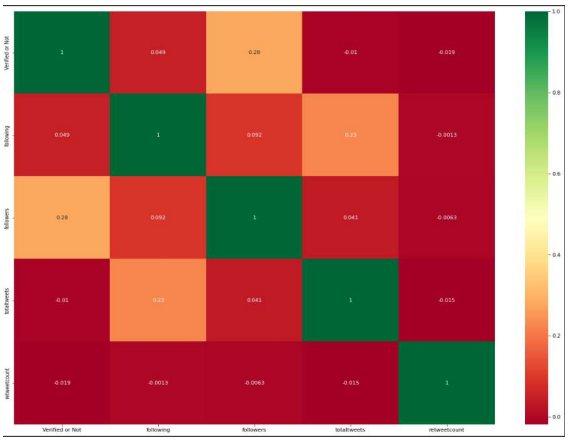

Figure 2. Correlation Heatmap of Data collected

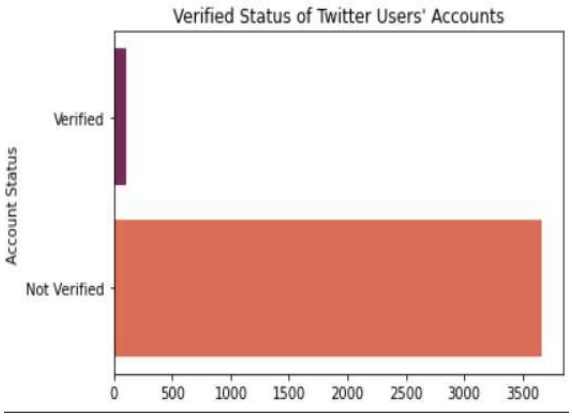

Figure 3. Verified Status Visualization

Figure 3. shows as well that the number of Verified accounts amounts only to 3% of the user data scrapped. With this, it can be concluded that the tweets posted were not necessarily from high-end influencers in this social media, and by default, not directly affecting the retweet counts, making the highest interactions with the tweets unrelated to who is tweeting it.

As for the sentiment generated, the analysis was split to two sub-questions:

- **What is the general sentiment?**
- **What is the most popular sentiment?**

The difference between the two is that we wanted to see the general sentiment of all the data first, then filter only the highest tweets interaction-wise and see the popular sentiment conveyed.

For the first sub-question, we performed a straightforward sentiment analysis on the tweets, after cleaning the data from any duplicates based on ***text*** attribute. After getting the results, we then performed further data cleaning on the texts themselves to get rid of any irrelevant components (links, hashtags, mentions, emojis …), and then compared the results obtained.

Figure 4. illustrates that after the cleaning of the texts, the sentiment analysis improved, even if by slight differences, but correctly classified tweets further when they were mistaken to convey a different sentiment.

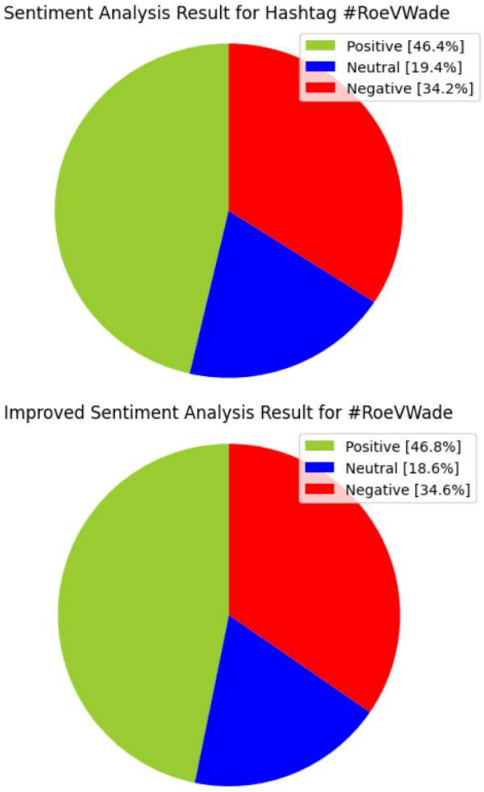

Figure 4. Visualization of sentiment analysis results before and after text cleaning

As for the second sub-question, we got the data of only the tweets with retweet counts higher than one hundred, obtaining a totality of 52 tweets. After performing sentiment analysis, Figure 5. was obtained, showing higher counts in the Positive sentiment conveyed.

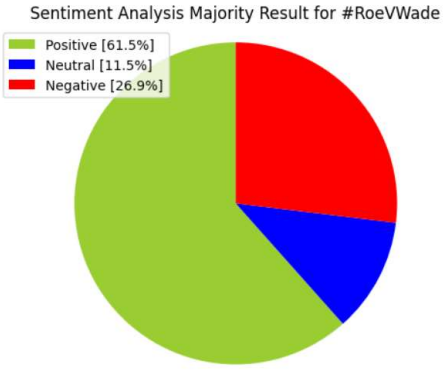

Figure 5. Sentiment analysis majority results

From all the pie charts, we can see the dominance of Positive followed by Negative sentiments. This is relative to what is the context of the tweets. For clearer clarification of what the machine thinks is pos/neg, word-frequency plots were generated to see what could be considered as positive and negative (Figure 6).

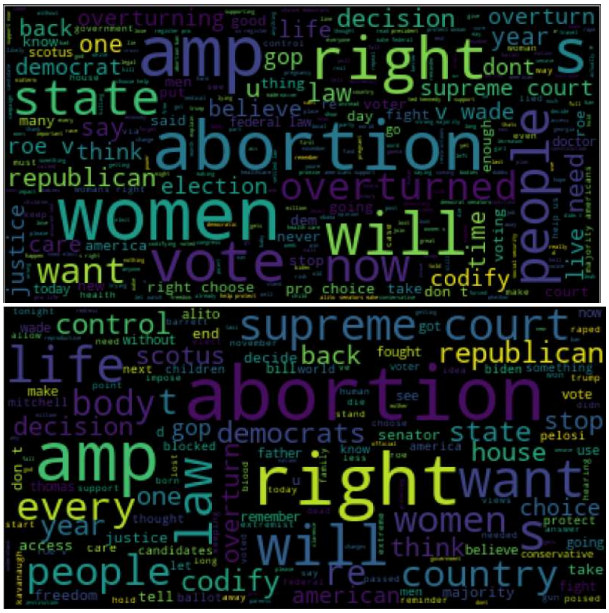

Figure 6. Word-frequency plots of general and major sentiments

From Figure 6., it is understandable that the sentiments conveyed would be both negative and positive in different senses. It could be positive in the sense that they are pushing more towards voting for overturning Dobbs v. Jackson itself and gaining back women's right to abortion, or that the conservative sides of the U.S are agreeing with such predicament. It could be negative in the sense that people are expressing how they do not agree with the overturning of Roe v. Wade and are taking it as obstructing women's right to abortion, or that they are expressing their view of limiting abortion for the rape cases or that the democrats are ruining the conservative values of the pro-

life believers. These interpretations were all mainly drawn from the words cited and prominent in Figure 6., although, it does not give a precise answer to RQ3, instead we only got possible speculations.

Conclusion

This study aimed to analyze and answer important questions around **#RoeVWade** that touched women's right to abortion in the U.S. This project came out with interesting insights and had varying depths of analysis to it, which helps observe the controversy around women's rights as a whole. Whether the perspective is pro or against the overturning, it is clear that women still have to face lots of limitations and control from different parties on multiple aspects of their freedom. This was one example from the country considered one of the most liberal, imagine what other laws are in place in other regions of the world, and how women have to live with such predicaments and face discomfort, unsafety, and oppression. Who is right and who is wrong? Who has the right and who does not? This can only be observed from a data scientist's perspective, and it is for the reader to interpret what the results mean to them.

Reference

1- Shi B, Chen J, Chen H, Lin W, Yang J, Chen Y, Wu C, Huang Z. Prediction of recurrent spontaneous abortion using evolutionary machine learning with joint self-adaptive sime mould algorithm. Comput Biol Med. 2022 Sep;148:105885. doi: 10.1016/j.compbiomed.2022.105885 . Epub 2022 Jul 26. PMID: 35930957.

2- Davidson L, Boland MR. Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. Brief Bioinform. 2021 Sep 2;22(5):bbaa369. doi: 10.1093/bib/bbaa369. PMID: 33406530; PMCID: PMC8424395.

3- Chen, Daniel L. and Kwan, Kristen and Quispe Ortiz, Luisa and Zamora Maass, Maria, Law and Norms: A Machine Learning Approach to Predicting Attitudes Towards Abortion (July 31, 2016). Available at SSRN: https://ssrn.com/abstract=2816659

4- Roe v. Wade - Wikipedia

5- How I Scrape and Analyse Twitter Networks [Case Study]