Abortion: Sentiments and Text Analysis from 2016-2020

Team 8 - Cho Xin Yi Jamie, Fathima S A Kabeer, Kwan Jia Ing Christine, Su Pei Xuan Lynn

ABSTRACT

Twitter and Reddit are microblogging platforms where people are free to post their thoughts and opinions about anything and everything. This paper centers on the study of the changes in sentiments on abortion over the last 5 years (2016 - 2020).

Using zero-shot classification and NLTK sentiment analysis, we conducted textual analysis on 55,000 posts to derive scores that indicate a post's positivity, negativity and neutrality score. Results from the 2 platforms show differing results where Twitter users have become more pro-choice while reddit users have remained more pro-life.

We discuss how our research can add value to other prior research, in particular "Shining the light on abortion: Drivers of online abortion searches across the United States in 2018". Additionally, we discuss how to further improve and develop our analysis in the future based on what we have gathered.

1   Introduction

This paper will study the sentiments and attitudes of how the public views abortion, and whether it has changed over the past 5 years (2016 to 2020). Abortion is defined as the termination of a pregnancy. As such, it is a rather controversial topic as it involves human life. Many individuals have been involved in this discussion on popular social media platforms such as Twitter and Reddit.

The two sides for the abortion debate fall under the terms "pro-life" and "pro-choice." People who are pro-choice believe that "everyone has the basic human right to decide when and whether to have children" [1], while people who are pro-life are those that oppose abortion. Abortion can be morally confusing as opinions vary based on one's own religion, political beliefs and ethics. Our research will study how changing times affect the general sentiments of abortion over the past 5 years.

1.1   Background Information

While many are concerned with the life of a fetus, other considerations such as welfare and financial capabilities of the woman are often overlooked. For instance, would it be more morally correct if the woman loses her life to conceive? Additionally, if the woman was unable to support the wellbeing of her child financially and emotionally, would the child be better off not being born? More often than not, the decision to abort is not black and white, which gives rise to many different schools of thought.
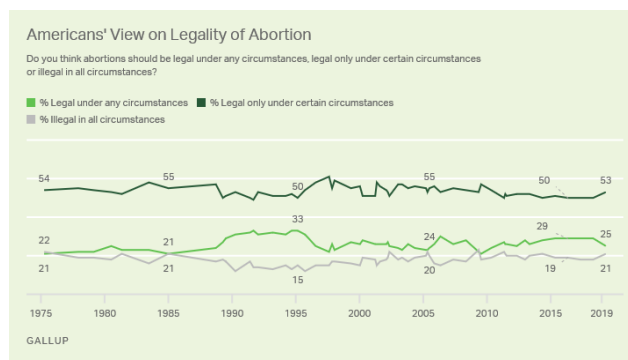

Figure 1: Views on Abortion in the US

A Gallup poll conducted in 2019 reported that 53% of American adults believe abortion should be allowed under specific circumstances while 21% think that it should be entirely illegal. Meanwhile, the remaining 26% favor abortion with no legal restrictions. Figure 1 shows a consistent trend for each of these 3 different categories of beliefs from 1975 to 2019 [2, Fig. 1].

Some recent movements that have gained traction revolve around feminism and human rights. These include "My Body My Choice" and the "Pro-Choice Movement." With many reputable organizations supporting abortion rights, such as the "National Abortion Federation" and "NARAL Pro-Choice America", it is no wonder more women are becoming more aware of the responsibilities of raising a child.

2   Key Concepts

We examined Twitter and Reddit posts between the years 2016 to 2020 to obtain textual data for our analysis.

Our corpus includes 40,000 tweets and 15,000 reddit posts. By drawing a large dataset, we will be able to generate more reliable and accurate estimates as a greater sample increases confidence, and precision of our results [3].

With the use of Natural Language Toolkit (NLTK) and zero-shot classification, we plan to analyze the emotional context of each post. NLTK provides us with sentiments on how positive, negative, or neutral a statement is. On the other hand, zero-shot classification calculates the relevance of a label in comparison to a tweet or reddit post. Additionally, we created bigrams for posts we scraped from Twitter and Reddit so we can understand why and what are the commonly talked about terms.

By plotting these figures on a time series graph with Microsoft Excel, we will be able to visualize changes in sentiment trends.

2.1 Hypothesis

With prior context from the background information found in section 1.1, we have arrived at two hypotheses. Appropriate steps will be carried out to determine the validity of our hypothesis.

Hypothesis 1: Attitudes and sentiments towards abortion have become more positive in the last five years for both Twitter and Reddit.

Hypothesis 2: Pro-choice sentiments have increased for both Twitter and Reddit.

3 Data Collection

3.1 Twitter

Upon preliminary searches on Twitter, we realised that keywords such as "abortion" and the phrase "my body, my choice" are commonly associated and relevant with our topic. Posts containing these terms were opinionated and varied greatly from one post to the other. Tweets were scraped using Snscrape with the query "abortion" and "my body, my choice.". Since Snscrape has a limitation of only scraping the most recent tweets, we scraped 5000 tweets each year from 2016 to 2020 for each of the 2 keywords. Attributes collected from the tweets include date, time, tweet ID, username and text. After scraping the tweets, we analysed 25,000 tweets per keyword using HuggingFace's zero-shot classification and NLTK Vader. Results will be discussed in the following sections.

3.2 Reddit

Using PRAW and PSAW's Push Shift API allowed us to retrieve posts from a particular subreddit within a given time frame. PSAW's limitation is that it does not allow for specific query searches. Hence, we searched for an appropriate subreddit by trying out different keywords. Eventually, we found that the subreddit - r/abortion was very relevant as it is an active community with over 21,000 users who post regularly. We retrieved posts from 2016 to 2020, generating a total of 15,000 reddit posts. Attributes collected from reddit posts include created_utc, number of comments, score, text body, subreddit, title, url and date created. Since the created_utc of the post is not in a human readable format, we created a new column in the dataframe called new_dates where we converted UTC timestamps into the YY-MM-DD format. We also merged content from the title and body columns so we can run the entirety of a post in NLTK and zero-shot.

4 Analysis Techniques

Apart from zero-shot Classification and NLTK Vader, we also skimmed through the raw textual data to gain insights on what people are posting. Additionally, bigram word clouds help us to further understand what is commonly talked about in the topic of Abortion.

4.1 Zero-shot Classification

Zero-shot classification from HuggingFace allows us to categorize data, in this case - Reddit and Twitter posts, using a set of labels. We have decided to use "pro-life" and 'pro-choice' as the labels since it clearly represents the 2 opposing sides in the debate of abortion. Zero-shot classification works by returning the probability scores for how relevant a piece of text is to a label. The higher the probability, the more significant that piece of text is. We then categorized each post by choosing the candidate label with the higher probability.

4.2 NLTK Vader

NLTK Vader produces the negativity, positivity and neutrality scores for each tweet and reddit post. Since Vader takes into account informal writing and slangs, we felt that this made it an accurate tool for analysing social media posts. It also provides a compound score, which standardizes the score from -1 to 1. -1 being the most negative and 1 being the most positive [4]. This would give us a clearer picture of how negative or positive a post is as compared to other posts.

After running data through the above-mentioned classification techniques, we proceed to visualize the results through Microsoft Excel graphs by plotting sentiment results against time to study the trends.

4.3 Bigrams

We decided to use bigrams analysis as it is more insightful as compared to unigrams. We made bigram word clouds from both Twitter and Reddit to further understand what are the commonly talked about terms regarding the topic of abortion.

5 Analysis Results

5.1 User Demographics

As both platforms originate from the US, it is no surprise that majority of the Twitter and Reddit demographic consists of Americans. A study finds that the US is the leading country with 73 million or 35.4% of Twitter users [5]. Similarly, Reddit's active users mainly comprises Americans, accounting for up to 48.9% of traffic [6]. The demographic statistics for other nationalities are pale in comparison. Therefore, for the purpose of this research, we will pivot our analysis around the US to determine plausible explanations for the analysis results.

5.2 US Politics

Based off related work regarding sentiments on abortion, we have gathered that abortion trends and sentiments tie in greatly with US politics. It is said that "pro-lifers are usually seen as conservative, often belonging to the Republican camp, while pro-choicers are

usually regarded as liberals who lean towards the Democrats" [7]. With that being said, we believe that the fluctuations found in the graphs below can be explained through significant events that occurred during political bouts over the past 5 years. The period of analysis (2016 to 2020) aligns with the presidency of Trump. Trump belongs to the Republican party and has taken significant action to protect life and human dignity. His voters, most of which are pro-life are in conformance with his cause and beliefs. Therefore, as one would expect, pro-lifers would not feel the need to become more vocal about their movement. Pro-choicers, on the other hand, would feel the need to spread the message about their cause, garner support and rally. This is evident in the Twitter graphs below. The Twitter graphs follow a trend in which pro-choice sentiments and mentions show a positive increase. Twitter showed that users leaned more towards the pro-life camp before switching over to pro-choice in the next few years.

5.3 Zero-shot Classification

5.3.1 Zero-shot classification accuracy

Figure 2: Zero-shot accuracy

40 reddit posts and 40 tweets were randomly selected to manually check for accuracy of zero-shot's classification's labelling results. The above image shows the results [Fig. 2]. Manual accuracy checks have been compiled into "accuracy check.csv". We felt that accuracy results of above 80% is good, therefore we proceeded with using this analysis method.

5.3.2 Twitter zero-shot results
Visualizations for Twitter and Reddit were generated using the data obtained from HuggingFace's zero-shot classification and NLTK. Based on the candidate labels (pro-life, pro-choice), the count of each class per year was chosen as the y axis.
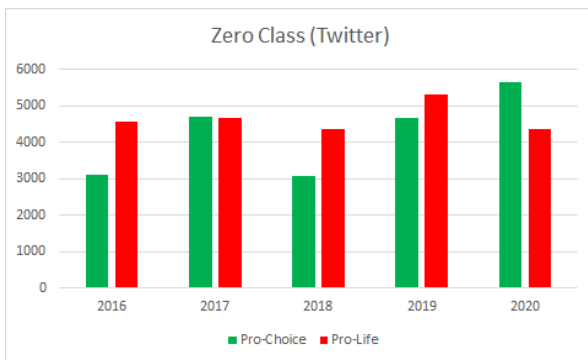

Figure 3: Zero-shot classification for Twitter (Bar)

As seen in the graph [Fig. 3], pro-life mentions continuously outnumbered pro-choice mentions up till 2019. In 2020, the pro-choice camp precedes pro-life. This change could be attributed to

the then upcoming US presidential election of Biden who supports and advocates the pro-choice movement. Users may have been expressing themselves in support of Biden's views on abortion.
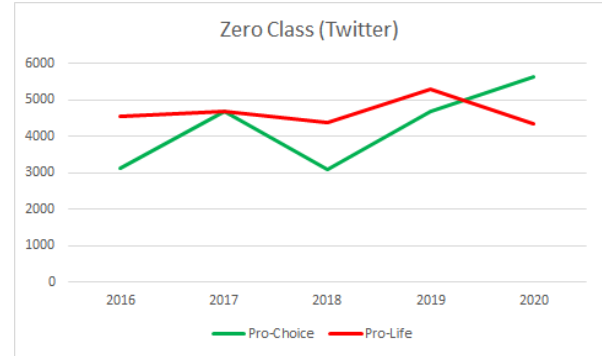

Figure 4: Zero-shot classification for Twitter (Line)

The above line chart [Fig. 4] is an alternative visualization of the change in sentiment towards abortion in 2020. This trend supports our hypothesis that supporters for pro-choice have increased on Twitter.

5.3.3 Reddit zero-shot results
As mentioned earlier, findings from Reddit do not tally with results from Twitter. Reddit's trend displays a consistent increase in pro-life mentions. Reddit, being mainly a discussion platform, may have more constraints as people are more conservative. This will explain why Reddit users support pro-life.
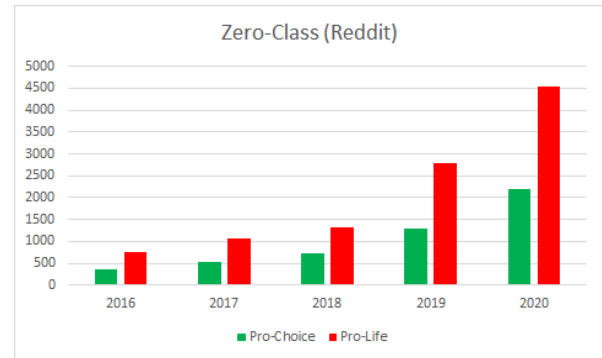

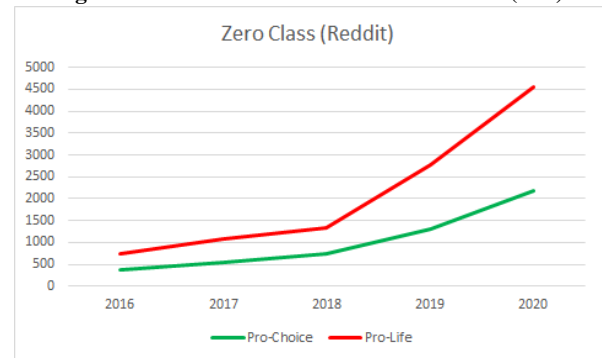**Figure 5: Zero-shot classification for Reddit (Bar)**

Figure 6: Zero-shot classification for Reddit (Line)

As seen the graphs above [Fig. 5-6], pro-life mentions continuously increase from 2018. It is also notable that pro-choice mentions show a gradual increase every year. This supports our hypothesis that supporters for pro-choice have increased on Reddit.

Graphs generated using data from Reddit show pro-life mentions have always been more popular from 2016 to 2020. A reasonable explanation for this could be the difference in purpose and user demographics between the two platforms. For instance, Twitter is more of a microblogging site while Reddit is discussion based. Twitter is also used to deliver fast and urgent messages [9]. Their tight word count of 140 characters urges users to keep information concise, thus making the sharing of information quick and easy. Also, Twitter has adopted the 'woke' culture. The term 'woke' refers to being alert towards prejudice, discrimination and other social issues. Therefore, it would make sense for Twitter to be the choice of platform for activism and advocacy. Thus, we conclude that Twitter, being more liberal as compared to Reddit, attracts more vocal people.

5.4 NLTK Vader

5.4.1 NLTK Vader accuracy

Twitter vader	0.675
Reddit vader	0.6

Figure 7: NLTK Vader accuracy

Similar to Zero-shot classification, we looked at 40 reddit posts and 40 tweets to check for accuracy of NLTK Vader. The above table shows the accuracy results [Fig. 7]. Although it is less accurate than zero-shot classification, we felt that the majority of it was still appropriately labelled, hence we carried on with the analysis.
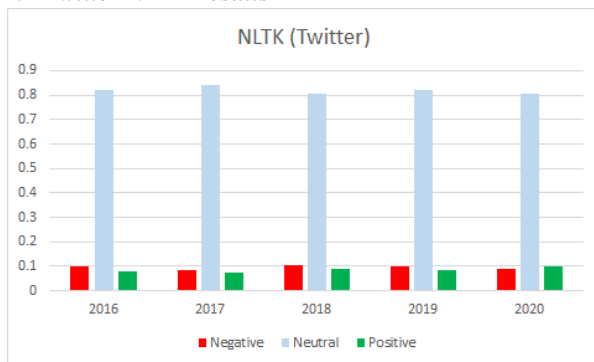
5.4.2 Twitter NLTK results

Figure 8: NLTK Sentiment Analysis for Twitter with neutral sentiments

The bar chart generated with neutral sentiments [Fig. 8] makes it difficult to compare positive and negative sentiments. Therefore, neutral sentiments from tweets and posts have been removed.
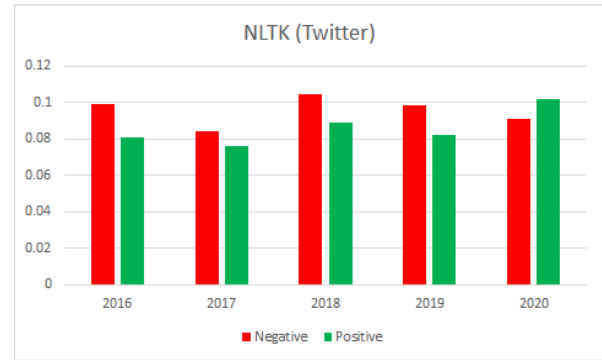

Figure 9: NLTK Sentiment Analysis for Twitter

By generating charts based on the average positive and negative value by year, it can be observed that for the first time in 2020, positive sentiments towards abortion overtook negative ones [Fig. 9]. It is an interesting observation that may be attributed to the then upcoming US presidential elections of one of the candidates, Biden, who supports pro-choice. As positive sentiments are seen to be increasing from 2019 to 2020, it does not fully support our hypothesis that attitudes and sentiments towards abortion have become more positive in the last 5 years on Twitter.
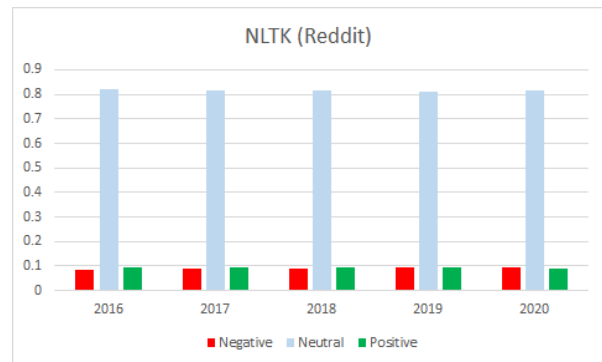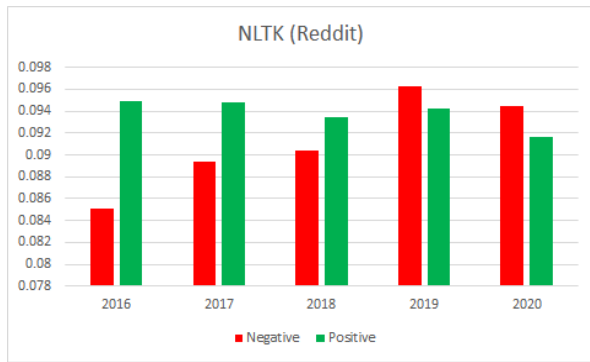
5.4.3 Reddit NLTK results

Figure 10: NLTK Sentiment Analysis for Reddit with neutral sentiments

Shown above [Fig. 10] is a bar chart comparing sentiments on Reddit. As neutral sentiments make up the bulk of our dataset, it will be removed so we can visualize trends between positive and negative sentiments.

Figure 11: NLTK Sentiment Analysis for Reddit

As seen in above [Fig. 11], posts show varying negative sentiments towards abortion. In 2016 to 2018, posts were found to display more positive sentiments while the opposite is seen from 2019 to 2020. In 2019, negative sentiments are at an all-time high, which is likely due to the fact that in 2019 the US state policies enacted to restrict abortion. Therefore, this trend could be indicative of people expressing negative feelings towards the ban. These findings do not support our hypothesis. An increase in positive sentiments towards abortion are not seen from 2016 to 2020.

5.5 Bigram Word Clouds

By generating bigram word clouds, we are able to understand which terms are commonly discussed on Twitter and Reddit. We have combined textual data from "abortion" and "my body, my choice" from Twitter into one word cloud. Additionally, we added the keywords used into stop words.

Figure 12: Twitter bigram word cloud

Figure 13: Reddit bigram word cloud

As seen from the Twitter word cloud [Fig. 12], more notable terms include "planned parenthood", "birth control", "pro life", and "get abortions." When cross checked with the raw dataset, it turns out that many people feel that women getting abortions are a personal choice. Additionally, many posts talked about making birth control more available so unwanted pregnancies can be avoided.

Regarding Reddit's word cloud [Fig. 13], words like "surgical abortion" and "medical abortion" as one of the most frequent bigrams. When we checked with the raw dataset, it appears to be that many people are posting about their experiences with medical and surgical abortion.

We conclude that people on Twitter talk more about their thoughts on abortion while Reddit has more people sharing their own encounters with abortion.

6 Past and Future Research

A past research titled "Shining the light on abortion: Drivers of online abortion searches across the United States in 2018," was carried out to find the main drivers for abortion searches on Google. The aim of this research was to derive what factors influence the traffic volume of abortion query searches [10]. One of the factors that was studied was whether areas with high abortion search volumes comprised of more pro-choice citizens. Our research techniques can provide them with a more solid understanding of the sentiments expressed on abortion via social media.

In the future, this study can be further improved by researching deeper into the demographics of individuals who support pro-life and pro-choice respectively. For instance, we can survey different age groups and genders to gain deeper insights on how these different demographics may affect one's opinion toward abortion. We can also study which political party a pro-life or pro-choice individual would support. This is especially the case in the US where politics play a huge role in one's mentality and opinions. Additionally, the context of each tweet or post can be considered. Some tweets and posts might not be appropriate as the user who posted it may not be talking about the topic seriously and hence making the results less accurate. For instance, sarcasm, word play and simply just throwing hate can skew results. By doing so, we can improve the accuracy of our work.

7 Related Work

A paper done by BioMed Central (BMC) Health Services reveals their findings on women's abortion experiences [8]. The method used is similar to ours, as they utilized keyword searches in multiple databases, such as Medline amongst other sources such as journals and articles to perform their study. Similarly, the research paper also focused on a fixed time period (1998-2007) to gather data on women's experiences. However, it was also stated that their sample size was small and participants self-volunteered, which might have

skewed the results since participants might have felt the need to answer in a certain manner.

The study was then split into two categories, experiences with medical abortion or surgical abortion. Medical abortion is more related to our research as it is more general and focuses on issues such as culture, role of the male in an abortion and remembrance of women's past abortions, all of which can be categorized under sentiments. Body autonomy, emotional experiences and moral values were also discussed in this paper. Women tend to feel more guilty going through a surgical abortion as compared to a medical one as the "prolonged nature of the procedure, including the sight of the fetus" induced more pain as compared to the medical approach.

Findings were conclusive where "political, ethical, social and legislative environments as well as health, economic and welfare systems" all played a part in the varied experiences of women's experiences during abortion. In fact, one conclusion made was that well-informed and supported women encounter positive results with abortion because they were aware of the responsibilities of becoming a mother.

With reference to our report, the study performed by BMC Health Services supports the pro-choice argument, implying that women should have full control of their body and that they should know whats best for their child. As such, pro-choice sentiments are improving.

8   Challenges and Limitations

During the data collection process, we made use of specific keywords to find posts that are relevant to our topic. This could be easily done for Twitter, but it was not the case for Reddit. PSAW only allowed us to scrape from a specific subreddit and not keyword queries. It is limiting because we were unable to check the content of all the reddit posts scrapped due to the sheer volume of data. Irrelevant posts might have slightly skewed the sentiment analysis.

For the purposes of understanding our results, we explained abortion sentiments by drawing similarities to US politics. Based on statistics, while it is true that majority of users who post on Twitter and Reddit are Americans, we cannot assume that all of them are Americans, or that all of them form their opinions on abortion because of the state of US politics alone.

Lastly, Snscrape only fetches the latest tweets. Due to the large number of posts on this topic, we tried to improve our dataset collection by retrieving 2000 tweets each year. However, this still causes limitations as tweets from the start of the year are not included in our dataset, with only the latest tweets of each year being represented for the entire year.

9   Conclusion

Data was collected from Twitter and Reddit over the time period of 2016 - 2020. The data was analysed using Zero-shot classification and NLTK sentiment analysis. In Zero-shot classification, we assigned labels "pro-life" and "pro-choice" to the Tweets and Reddit posts. While for NLTK sentiment analysis, we looked at the positive, negative and neutral sentiments of the posts. We plotted both Zero-shot classification and NLTK sentiment results on a time series chart to observe the changes in sentiments of pro-choicers and pro-lifers. Additionally, user demographics of Twitter and Reddit played a huge role in our data analysis as the majority of users are from the US. Through our analysis we observed that the tweets and Reddit posts were significantly affected by the US presidential election in 2020. The topic of abortion is controversial and therefore, widely talked about in major political campaigns.

In conclusion, hypothesis 1 is inaccurate. While it is true that sentiments on abortion have become more positive on Twitter, this is not the case for Reddit. Hypothesis 2 is also false since Reddit had a surge of pro-life supporters. Twitter on the other hand, ultimately had more pro-choice believers by the end of the time period which we attributed to political campaigns held during US presidential elections in 2020.

REFERENCES

[1] Miriam, "Can you explain what pro-choice means and pro-life means? ," *Planned Parenthood*, 16-Oct-2019. [Online]. Available: https://www.plannedparenthood.org/learn/teens/ask-experts/can-you-explain-what-pro-choice-means-and-pro-life-means-im-supposed-to-do-it-for-a-class-thanks. [Accessed: 10-Nov-2021].

[2] L. Saad, "Majority in U.S. still want abortion legal, with limits," *Gallup.com*, 13-Aug-2021. [Online]. Available: https://news.gallup.com/poll/259061/majority-abortion-legal-limits.aspx. [Accessed: 10-Nov-2021].

[3] S. Littler, "The importance and effect of sample size," *Select Statistical Consultants*, 24-Aug-2021. [Online]. Available: https://select-statistics.co.uk/blog/importance-effect-sample-size/. [Accessed: 10-Nov-2021].

[4] J. Burchell, "Using vader to handle sentiment analysis with social media text," *Standard error Full Atom*, 08-Apr-2017. [Online]. Available: https://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html. [Accessed: 10-Nov-2021].

[5] Statista Research Department, "Twitter: Most users by country," *Statista*, 07-Sep-2021. [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. [Accessed: 10-Nov-2021].

[6] J. Clement, "Reddit: Traffic by country," *Statista*, 09-Nov-2021. [Online]. Available: https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/. [Accessed: 10-Nov-2021].

[7] S. Varrella, "United States - pro-choice and pro-life supporters 2020," *Statista*, 30-Sep-2021. [Online]. Available: https://www.statista.com/statistics/225975/share-of-americans-who-are-pro-life-or-pro-choice/#statisticContainer. [Accessed: 10-Nov-2021].

[8] M. L. S. Lie, S. C. Robson, and C. R. May, "Experiences of abortion: A narrative review of qualitative studies," *BMC Health Services Research*, vol. 8, no. 1, Jul. 2008.

[9] "Ask any difference," *Ask Any Difference*, 05-Aug-2021. [Online]. Available: https://askanydifference.com/difference-between-reddit-and-twitte/. [Accessed: 10-Nov-2021].

[10] S. Guendelman, E. Yon, E. Pleasants, A. Hubbard, and N. Prata, "Shining the light on abortion: Drivers of online abortion searches across the United States in 2018," *PLOS ONE*, vol. 15, no. 5, May 2020.