#MeToo Movement Analysis: Sentiment, Themes, and Impact

Introduction

Social Media such as Twitter plays a vital role in the globalization of issues and acts as a catalyst for uniting people for a cause. The #MeToo movement is a prime example of this. Twitter played a crucial role in providing a platform for survivors of sexual assault and harassment to speak up about their experiences as well as finding community and empowerment in others as well as themselves, inspiring many other survivors to share their stories. This has allowed voices that had been silenced for too long to be heard by not just the people they knew but the entire world.

Analyzing this movement will help us grasp the true essence of this movement and what it stands for. The sheer volume of tweets and the diversity of voices revealed the extent of the problem. It provided a data-rich environment for researchers and activists to study patterns, trends, and the public's reactions, shedding light on the nuances of the movement.

The #MeToo movement lacks a thorough and systematic analysis of its multifaceted dimensions. There is an imperative to conduct an in-depth evaluation that encompasses public reactions, assesses the movement's influence on incidents of sexual harassment and assault, and monitors the engagement of key figures within the movement. This analysis seeks to address the need for a comprehensive understanding of the movement's impact, challenges, and key contributors.

Table of Contents

Understanding the Problem Statement

The problem we aim to address is to conduct sentiment analysis on #MeToo movement twitter data to evaluate the reaction of the public towards the movement as well as to gauge the media's portrayal of the movement, to identify the key issues in the movement that needs addressing, and also to monitor the involvement of key individuals.

About the Dataset

- The dataset has been obtained from data.world. It was created by Brett Turner in 2017.
- The dataset includes text content of tweets posted between November 29th and December 25th, 2017.
- Each tweet is associated with metadata, including the timestamp, user information (username), and engagement metrics.
- The data may contain multimedia elements such as links and hashtags shared in tweets.
- Dataset is completely raw data scraped from Twitter by the author with no preprocessing done.

# First Look at the Dataset

Our data comprises 3,93,135 rows and 17 columns. The columns are:

- 'Unnamed: 0': the inherent index of the dataframe
- 'text': the actual text of the tweet
- 'favorited': if the account scraping the tweet has liked the tweet
- 'favoriteCount': the number of likes the tweet has received
- 'replyToSN': the account to which the tweet is being replied to (null if the tweet is original)
- 'created': the date and time the tweet was created
- 'truncated': if the tweet was truncated during the recording process
- 'replyToSID': the account ID to which the tweet is being replied to
- 'Id': the tweet ID
- 'replyToUID': the user ID of the reply tweet
- 'statusSource': the device from which the tweet has been sent from
- 'screenName': the username of the tweet creator
- 'retweetCount': the number of times the tweet has been retweeted
- 'isRetweet': true if the tweet has been retweeted even once
- 'Retweeted': true if the tweet has been retweeted by the account scraping it
- 'longitude': the longitude of location where the tweet was created
- 'latitude': the latitude of location where the tweet was created
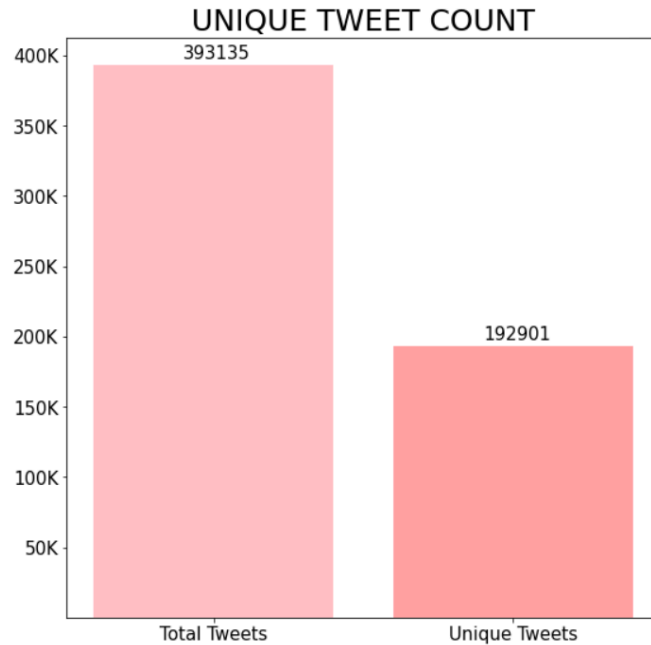
UNIQUE TWEET COUNT

This bar graph shows the total number of tweets and the number of unique tweets. It's natural for a tweet to be recorded twice during the scraping process especially when it has been scrapped in batches. We have dropped the duplicated tweets and preserved the one where the retweet count is the highest to preserve engagement information. Further graphs are plotted on a dataset free of duplicate tweets.

After deletion of duplicates, we were left with 1,92,901 rows.
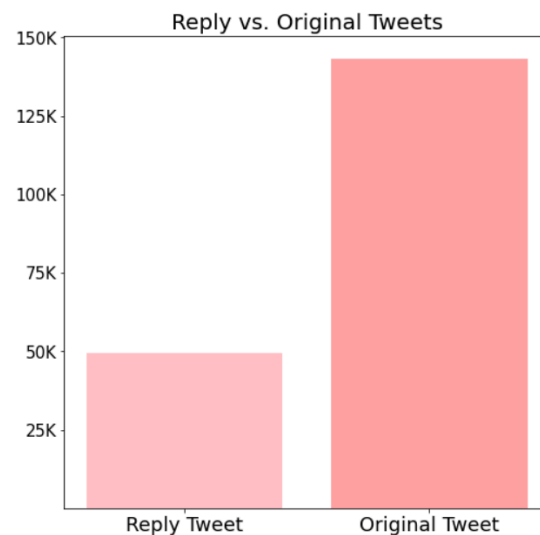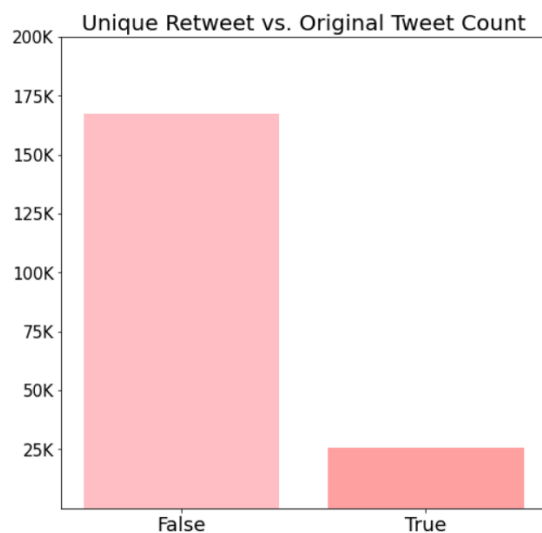
These bar charts compare the original tweets with tweets that are either replies or retweets of other tweets. Since the number of original tweets is higher in both the cases, we infer that our

data is rich and contains variety. The number of tweets that are replies to other tweets being less is a plus point because the tweets being original won't lack context.
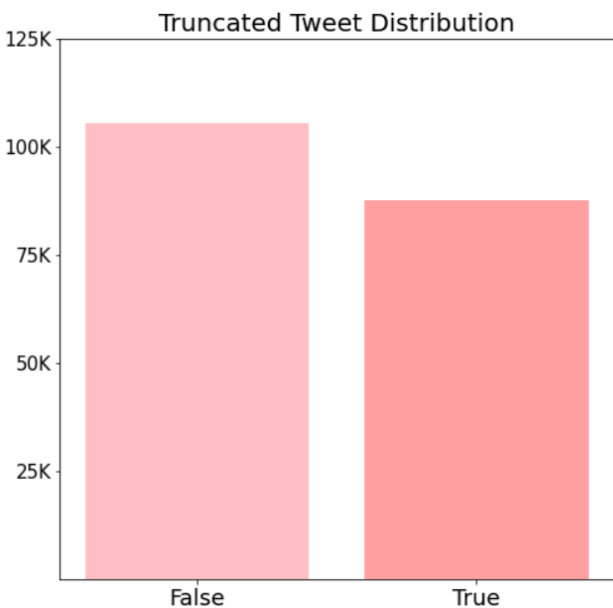
This chart visualizes the number of tweets that were or were not truncated during the recording process. The amount differs slightly between the two categories. Since the majority lies in the 'False' category, i.e, they are not truncated, we can be assured that our analysis will provide close to accurate results.
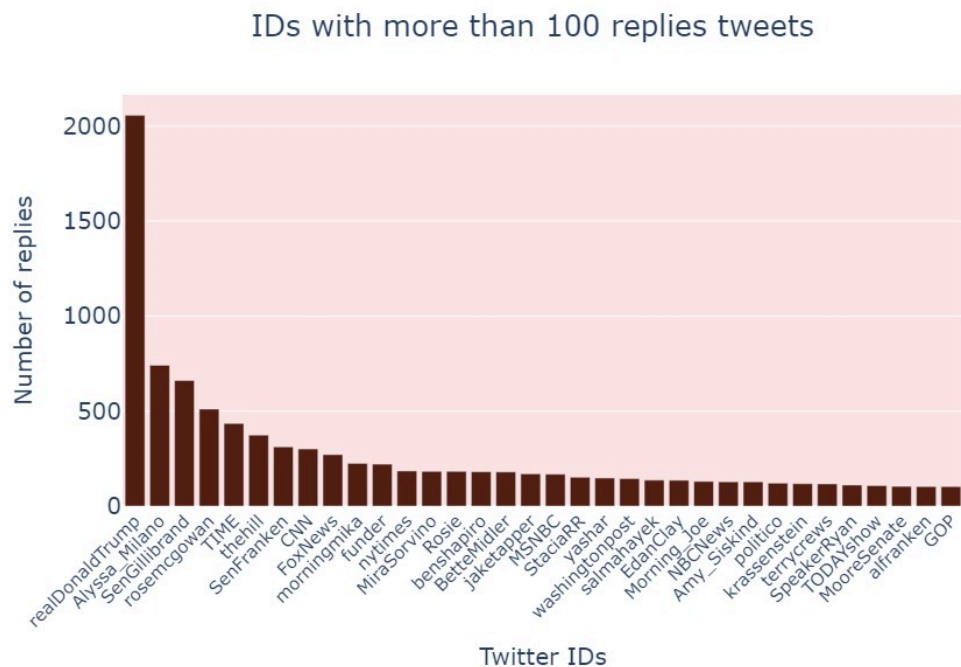
This chart visualizes the 'replyToSN' column. It indicates users to which maximum replies have been directed too. We can see that public figures like Donald Trump, Alyssa Milano, Senator Gillibrand were involved in this movement along with news channels like TIME, CNN, NBCNews, FoxNews etc.

Preprocessing

The text column in the dataframe that contained the tweets were littered with links to the actual tweet, usernames and hashtags. Natural text also contains punctuations and contractions. These anomalies if not eliminated at this stage impact the EDA, sentiment analysis and topic modeling negatively.

The steps of preprocessing that we followed:

1. Removal of html links: Achieved by using the re library of python and specifying a pattern that matches links

2. Removal of usernames: Achieved by using re library by specifying a pattern that matches the usernames

3. Removal of hashtags: Achieved by using re library and specifying a pattern that matches hashtags. The tweet text with the hashtag was preserved in a different column for further analysis. A new column was made for tweets without the hashtags.

4. Removal of 'RT': These were at the beginning of retweets and were necessary to eliminate.

5. Removal of other artifacts like underscore, quotation marks etc: Achieved by using re library

6. Conversion of text into lowercase: To make all the tweets uniform.

7. Expanding contraction: It was necessary to expand all contractions such as "i'm", "don't" into "I am" and "do not" for effective lemmatization. It was achieved by using the contractions library.

8. Removing Punctuation: Since punctuation do not provide any meaningful information in the scope of our analysis, we removed them using the string library of python.

9. Removing stopwords: Stopwords create noise and unnecessarily increase the computing time. We eliminated them using the NLTK library by loading the list of stopwords in the English language. A separate column was created for tweets that were devoid of stopwords.

10. Lemmatization: It is necessary to reduce the word into their root form for tasks such as topic modeling and sentiment analysis, hence we lemmatized the tweets using the medium sized pre-trained model from the spacy library. The lemmatized tweets were stored in a new column.

	text	text_without_hashtag	expanded_tweet	punctuation_removed	tokenized_tweets	stopwords_removed	lemmatized_tweets
0	american harem.. #metoo	american harem..	american harem..	american harem	['american', 'harem']	american harem	american harem
1	why have you guys not resigned yet? liberal hy...	why have you guys not resigned yet? liberal hy...	why have you guys not resigned yet? liberal hy...	why have you guys not resigned yet liberal hyp...	['why', 'have', 'you', 'guys', 'not', 'resigne...	guys resigned yet liberal hypocrisy	guy resign yet liberal hypocrisy
2	watched megan kelly ask joe keery this a.m. if...	watched megan kelly ask joe keery this a.m. if...	watched megan kelly ask joe keery this a.m. if...	watched megan kelly ask joe keery this am if s...	['watched', 'megan', 'kelly', 'ask', 'joe', 'k...	watched megan kelly ask joe keery rub fingers ...	watch megan kelly ask joe keery rub finger hai...
3	women have been talking about this crap the en...	women have been talking about this crap the en...	women have been talking about this crap the en...	women have been talking about this crap the en...	['women', 'have', 'been', 'talking', 'about', ...	women talking crap entire time finally someone...	woman talk crap entire time finally someone li...
4	. please speak to this sexual assault by duri...	. please speak to this sexual assault by duri...	. please speak to this sexual assault by duri...	please speak to this sexual assault by durin...	['please', 'speak', 'to', 'this', 'sexual', 'a...	please speak sexual assault interview	please speak sexual assault interview

A snapshot of the data frame after the preprocessing. All the other columns are still intact.

Exploratory Data Analysis

We have visualized the most popular tweets to see how it affects the discussions at that point in time. This analysis helps us understand when the discussion is peaking in volume and leads us to analyze what is being discussed then.

Timeline of the 1000 Most Popular Tweets
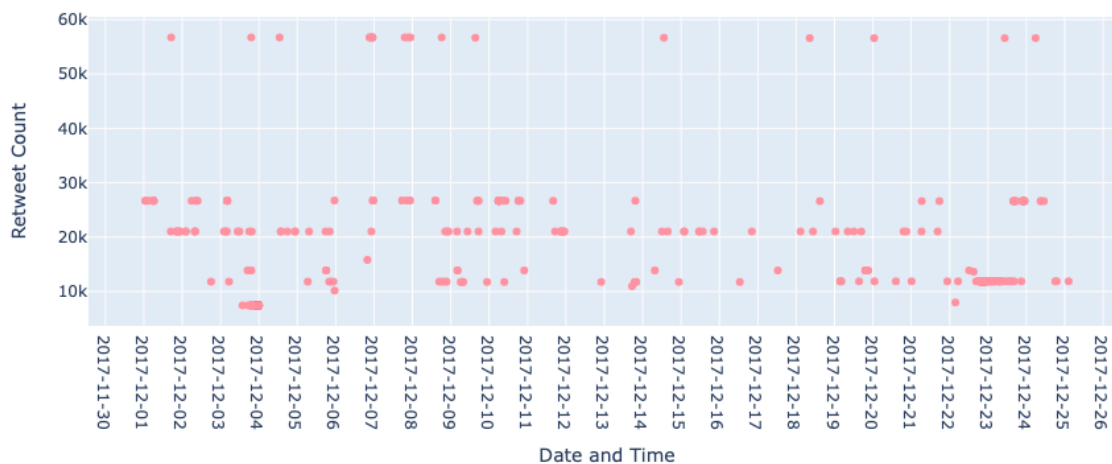
We found the people with the most tweets in 3 categories: Overall, Most Retweeted and Most Popular.

```
People with the most tweets:
DulleyTopBooks: 790 Tweets
evry2min: 330 Tweets
pitbulparalegal: 251 Tweets
dubie1151: 227 Tweets
DeWayne_Watts: 227 Tweets
PaganMediaBites: 193 Tweets
Auroraknite: 191 Tweets
sillymickel: 175 Tweets
PjjPjoje: 169 Tweets
DaShanneStokes: 168 Tweets
```

Overall: using the entire data we have checked the number of tweets per user to infer the most active people in the movement.

- Notable names in this list include DaShanne Stokes, an author, sociologist, public speaker who speaks on politics, culture, and civil rights and is known as a progressive activist.

```
People with the most tweets in top 10000 most retweeted Tweets:
DulleyTopBooks: 501 Tweets
DaShanneStokes: 168 Tweets
evry2min: 149 Tweets
dubie1151: 146 Tweets
RestlessNews: 136 Tweets
cindygallop: 128 Tweets
Blackamazon: 106 Tweets
ILoveBernie1: 96 Tweets
PjjPjoje: 90 Tweets
MediaShrink: 88 Tweets
```

Most Retweeted: Finding the users with most retweeted posts helps us understand the main people driving the discussion and shaping the public sentiment.

- Lucinda Lee Gallop, an English advertising consultant, founder and former chair of the US branch of advertising firm Bartle Bogle Hegarty, and founder of the IfWeRanTheWorld and MakeLoveNotPorn was a major contributor to the discussions as well.

```
People with the most tweets in the list oftop 10000 most popular Tweets:
funder: 12 Tweets
Alyssa_Milano: 3 Tweets
iamLucretiaMott: 3 Tweets
StashRiders: 2 Tweets
nelmadman: 2 Tweets
loveforlecki: 2 Tweets
lucilleaustero2: 2 Tweets
CherokeeNative3: 2 Tweets
ohsoveryzooey: 2 Tweets
elleduncan: 2 Tweets
```

Most popular: Here using the most retweeted and most liked tweets, we have found the key players of the movement who have made the greatest impact.

- Here we can also infer some names that are known such as Alyssa Milano, whose tweet was the one that sparked the movement to life.
- Another prominent name who was on the list is Elle Duncan, an american sports anchor for ESPN who expressed her frustration during an appearance on The Undefeated's podcast The Intersection where she spoke about the Washington Football Sex Scandal saying "No one stands up for Women".

Next, we got into the basic details of this movement — words. This movement was built on words so it is essential to see what kind of words were popular in this movement and the reason behind their popularity.

50 Most Popular Words
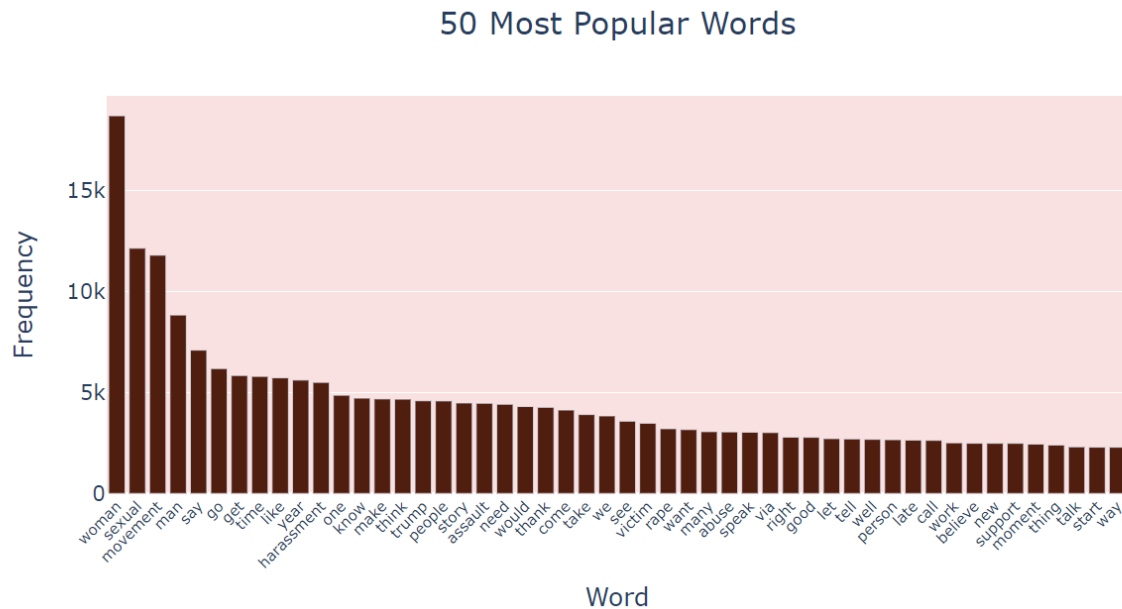
This chart illustrates the top 50 words. From this we picked out words that were worth exploring and printed 10 most liked tweets containing that word to understand the different contexts in which that word was used. We relied on 'likes' as a user-engagement metric because people like a tweet only if they agree with it. In this way we can only consider 10 tweets in practice but understand the opinion and thoughts of thousands of people i.e the people who agreed or liked with that tweet.

i. **'Assault'**

The tweets containing the word 'assault' talk about sexual assault and sexual assault allegations. They also talk about awareness and empathy. People are supporting sexual assault survivors. Trump has 19 sexual assault allegations against him. People want him to resign. These tweets also indicate a shift regarding sexual assault allegations in the Congress, since they are demanding investigation into the allegations against trump. The members of Congress want him to resign. This tells us that the #MeToo movement sparked a change in the way sexual assault allegations are handled at a political level.

ii. **'Rape'**

In the top tweets containing the word 'rape', the term 'rape culture' has been mentioned quite a few times. Rape culture is the name given to a society where sexual assault is trivialized or normalized. These tweets are addressing and acknowledging the silent pain of people living in

such a society. It also highlights the grim truth of how the government decides if a woman was raped or not. There are mixed opinions about this since some are against the idea that the society we live in today is patriarchial and refuse to accept that rape culture exists. Survivors speaking about their experiences and name dropping where celebrities like Danny Masterson, Geraldo Rivera were mentioned.

Geraldo River later admitted to assaulting Bette Milder in the bathroom before an interview. This movement brought the good outcome of Danny Masterson receiving a 30-year prison sentence after he was found guilty of two rape crimes after the survivor exposed him during the #MeToo movement.

iii. **'Speak'**
The word 'speak' was used in two different contexts. People are speaking in support of the abusers and are talking about ''speaking up' against abusers. The name Paul Ryan comes up who is in the same party as Donald Trump and hence the natural inclination to speak in favor of Donald Trump. Later as Trump's case worsened, Paul Ryan did realize that Trump was unfit to be the nominee of the Republican party.

iv. **'Support'**
The tweets containing the word 'support' have a relatively positive outlook. They praise Tarana Burke, the founder of the #MeToo movement. There are also tweets supporting the abusers.
These tweets also mentioned Juanita Broaddrick, America's former Nursing Home Administrator who spoke up after years of being raped by President Bill Clinton on April 25, 1978. This shows how the #MeToo movement enabled women to speak up after years of being assaulted. It provided them a platform.

Shiori Ito, a Japanese journalist and filmmaker went public in 2017 during the #MeToo movement with allegations that Noriyuki Yamaguchi, a former Washington bureau chief for the TBS network with close ties to the prime minister, Shinzo Abe, had raped her two years earlier. It was a class of rape called rape date where she was raped while she was under the influence of alcohol and in a non-consenting state. She ended up winning the lawsuit (compensated $30,000 in damages) and her action was meaningful to women all around the world. Again, it was the #MeToo movement that brought out these victories.

After analyzing unigrams, we moved on to analyzing bigrams.

Top 50 bigrams
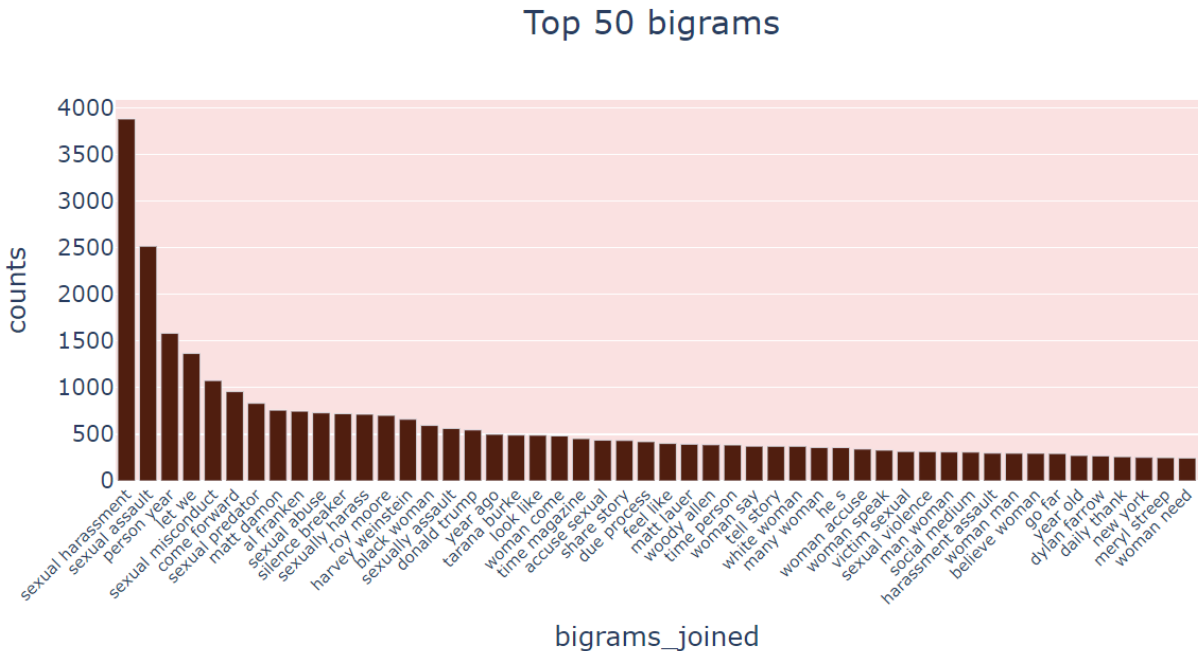
The bigrams that we chose to explore further were:

i. 'Black Woman'
The tweets that contained the bigram 'black woman' highlight an issue in the movement. Tarana Burke founded this movement in 2006 and yet she wasn't named 'person of the year' by the TIME Magazine. They blame the TIME magazine suggesting that just because she was a black woman the hashtag itself was named the person of the year rather than the founder. The tweets protested against the fact that she was not on the cover page either. These tweets highlight the issue of racism which spilled into this movement too. The #MeToo-s of black women were given less importance or undermined in comparison to those of the white women. These tweets also explain the occurrence of other bigrams and trigrams like 'time magazine', 'time magazine person', 'tarana burke', 'time person', '2017 person year'.

ii. 'Matt Damon'
Matt Damon, an American actor and comedian was criticized by the participants of the #MeToo movement for voicing his opinions. He was hated for saying that there was a "culture of outrage and injury," and that "there's a spectrum of behavior" that needed to be considered. "There's a difference between patting someone on the butt and rape or child molestation," he had said. This was aimed at various actresses like Alyssa Milano, Minnie Driver, and Evan Rachel Wood who made their feelings known on twitter. He was criticized by the participating women called "tone deaf".

Moving forward, we analyzed the trigrams too. Most of them were repetitions of the unigrams and bigrams.


Top 40 trigrams
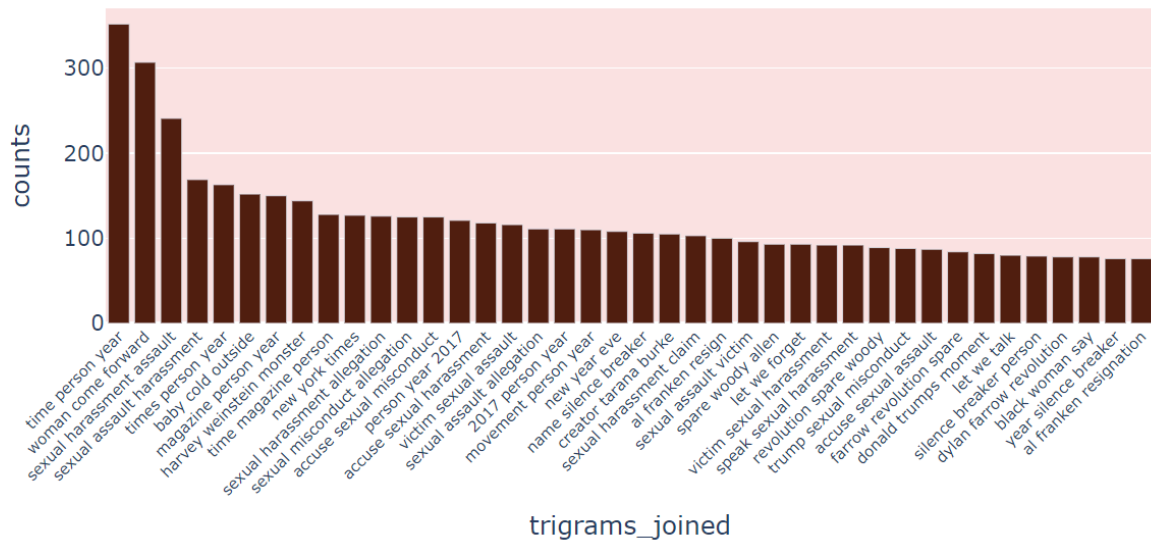
Here, the trigram 'baby cold outside' stood out to us so we dug deeper and extracted tweets that contained those terms. We found that 'Baby It's Cold Outside' is a famous Christmas song composed in 1944 always played during the holiday season. The tweets are questioning if the song is appropriate to be played in the #MeToo era. It's being called a sexual predator anthem.

The lyrics are written in the form of a conversation that goes back and forth between a man and woman where the woman wants to leave but the man insists she stay. He dismisses every excuse that she comes up with.

This song became problematic in the #MeToo era because it was made in an era when consent was not given much importance but as the movement gained momentum, the importance of consent became blatant and many critiqued people for defending the song. This song trended during the movement and due to people's negative opinion on it, it got banned from being played on major radio stations.

Top phrases
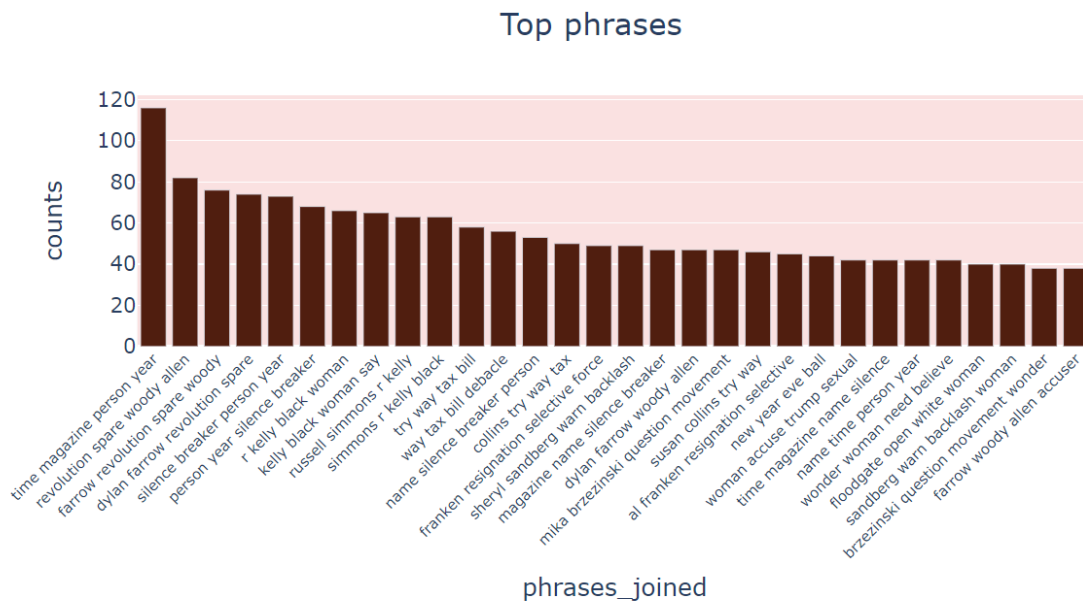
Next, we plotted a chart to visualize the top 100 hashtags to understand what other topics gained relevance in the movement.

Top 50 hashtags
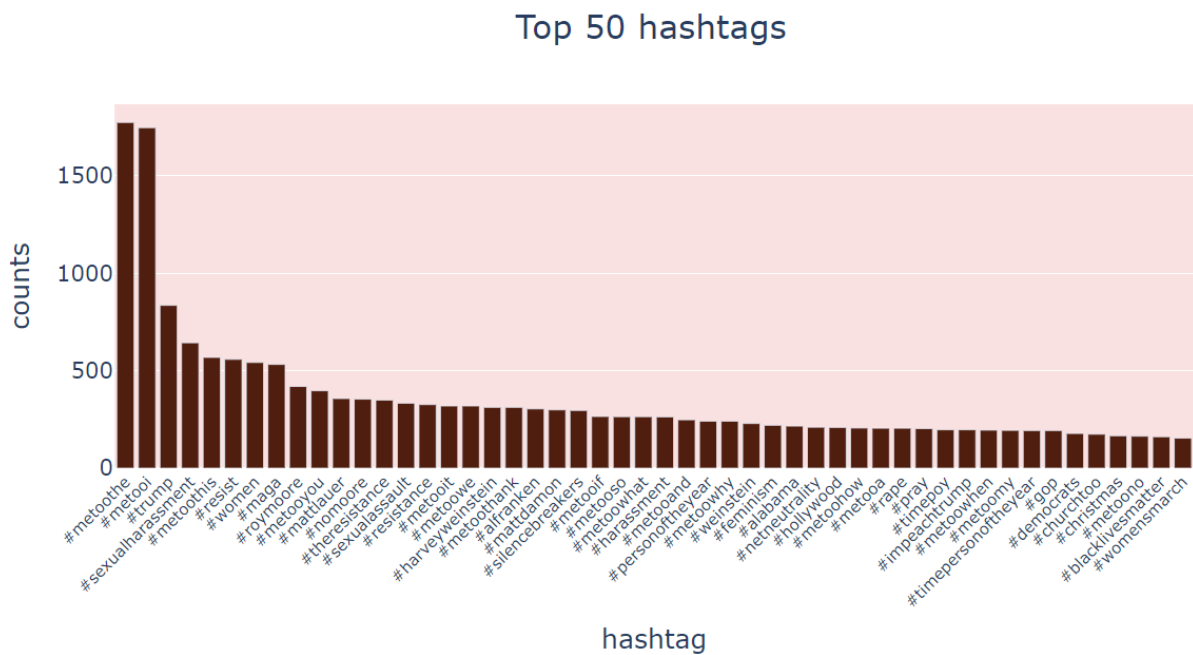
This chart is void of the #MeToo hashtag itself because it skewed the bar graph.

Next 50 hashtags
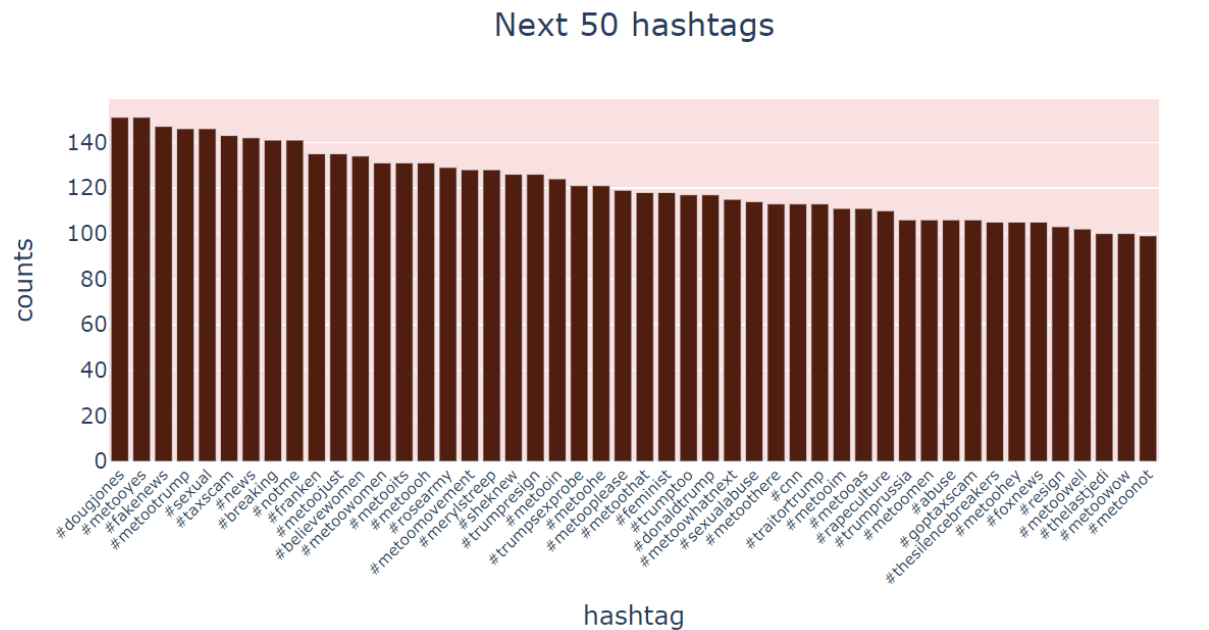
Sentiment Analysis

#MeToo movement is a global phenomenon that played a very important role in addressing the issues related to sexual harassment and assault. Sentiment Analysis plays an important role in understanding the sentiment of the public towards the movement, the key individuals involved, analyzing the trends of the movement and the sentiment across different geographic regions.

We have used VADER (Valence Aware Dictionary for Sentiment Reasoning) for analyzing the sentiments. After the initial preprocessing, it is necessary to do the preprocessing necessary for the respective objective.

In this particular objective, the preprocessing done is dropping duplicate tweets and only retaining the ones with highest retweet Count. The next step was to remove the tweets with two words or less because those don't really affect the sentiment.
The next step after this is to categorize the tweets as positive or negative or neutral based on the compound sentiment scores.

	punctuation_removed	sentiment_label
0	reminder that if a woman did not post it does...	Negative
1	i have decided its time to tell my story	Neutral
2	for my wife for my daughters for all womeni st...	Neutral
3	for those carrying their with them silently y...	Positive
4	at the very same time that we are asking frank...	Negative

Further after the labeling of tweets as positive or negative or neutral we visualize the sentiments.
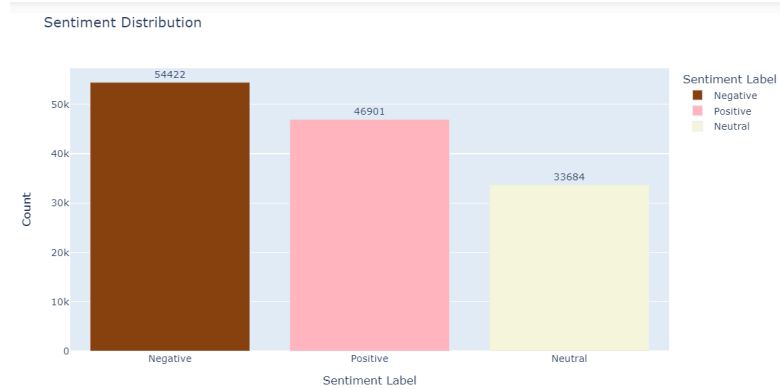
Here it is noticed that the negative sentiment has a substantial presence which indicates the negative nature of the sexual harassment and abuse. Positive sentiment was also represented which indicates support for the movement after the initial negative reaction to it. Neutral sentiments were also observed which provides facts without any emotional connect.

After the sentiment analysis of the tweet, we analyze the sentiment of the top 50 popular tweets based on the favouriteCount.
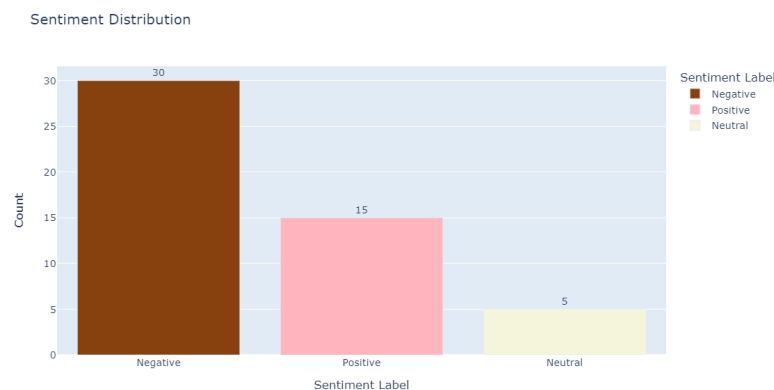

Through the analysis of the sentiments of the top 50 tweets, it is observed that the major sentiment is negative(30 tweets out of 50) followed by negative and then neutral sentiment.

Next step is to explore how sentiments change over time from 29 November 2017 to 25 December 2017 day wise. It was observed that the initial reaction to the #MeToo movement was negative. This was due to the skepticism, backlash and the fear of false accusations. But after a few days, particularly 6 December 2017 there was a noticeable shift towards positive sentiment due to release of the Time Magazine where #MeToo movement was named the person of the year. On the cover of the Time magazine were the people behind the movement "The Silence Breakers" which features the people who have spoken out. The positive sentiment lasts for a few days and then there is a shift towards the negative sentiment.
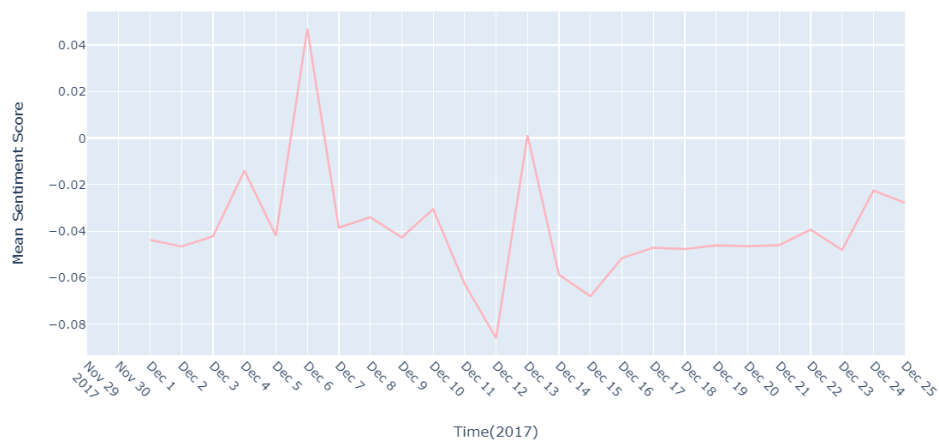
Next step is to examine the sentiment of tweets that are directed at well-known people. Top 50 people were chosen who were most replied to and the sentiment of those tweets towards those people was analyzed.
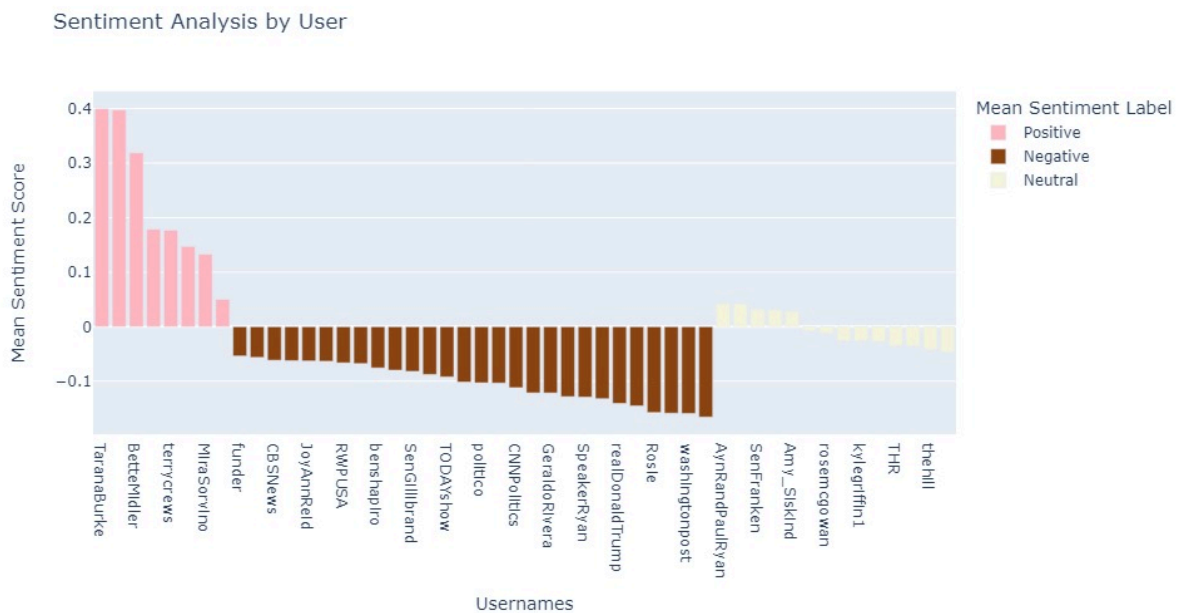

Donald Trump received a major negative sentiment directed at him because numerous allegations of sexual assault thrown at him as well as dismissive attitude toward these allegations. Tarana Burkewas directed a lot of positive sentiments because of her role in founding the #MeToo movement and also empowering the survivors.

Last objective was to visualize the sentiments of tweets across different geographic locations using the geopy library. Geopy library gives the address(country, state etc.,) from the latitude and longitude coordinates.

	Country	Mean Sentiment Score
0	Aruba	0.296000
1	Australia	0.000000
2	België / Belgique / Belgien	0.000000
3	Brasil	0.458800
4	Canada	-0.008017
5	Deutschland	0.000000
6	France	0.000000
7	India	-0.143467
8	Nederland	0.700300
9	Nigeria	0.765000
10	Polska	-0.421500
11	Singapore	0.000000
12	United Kingdom	0.219167
13	United States	0.112846

After calculating the mean sentiment score, the visualization was done to see the sentiment of tweets across the geographic locations.
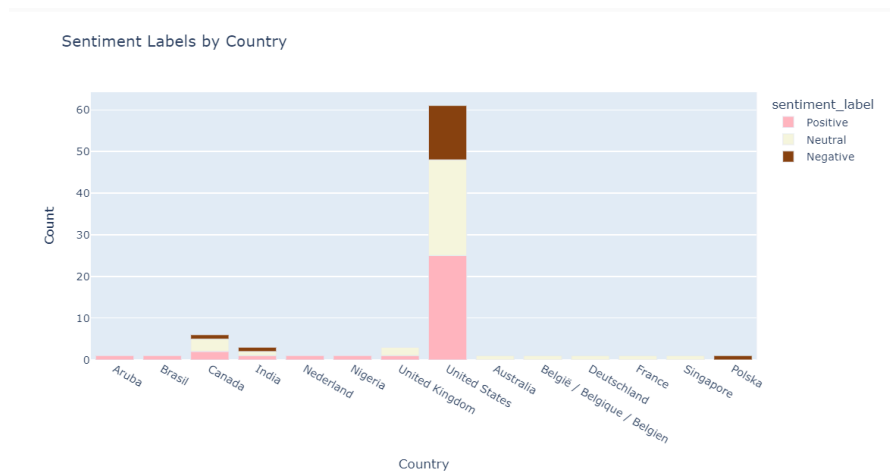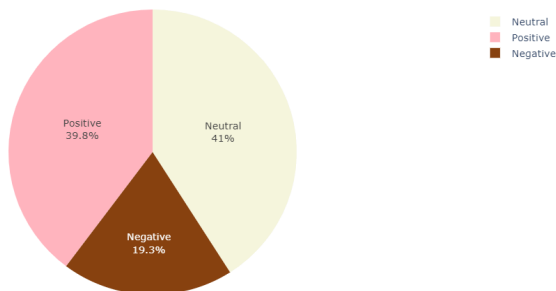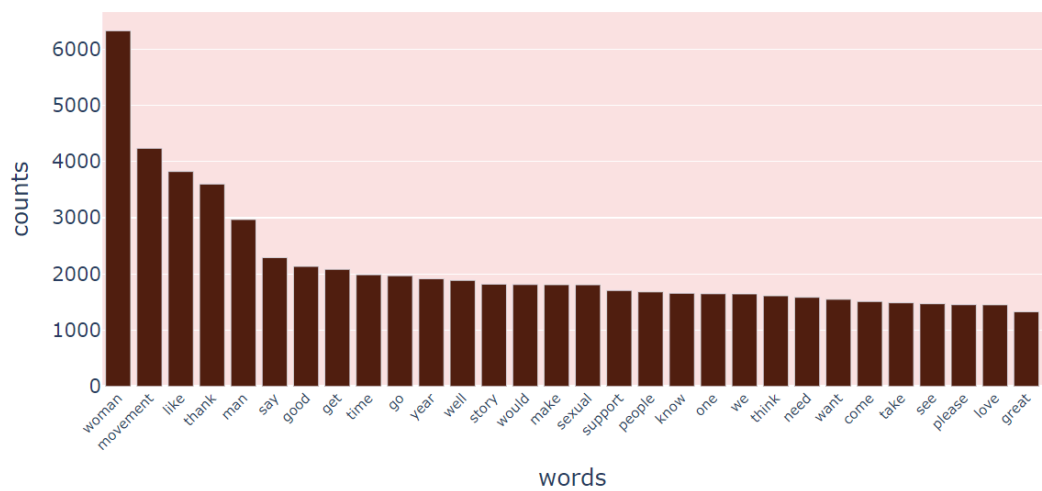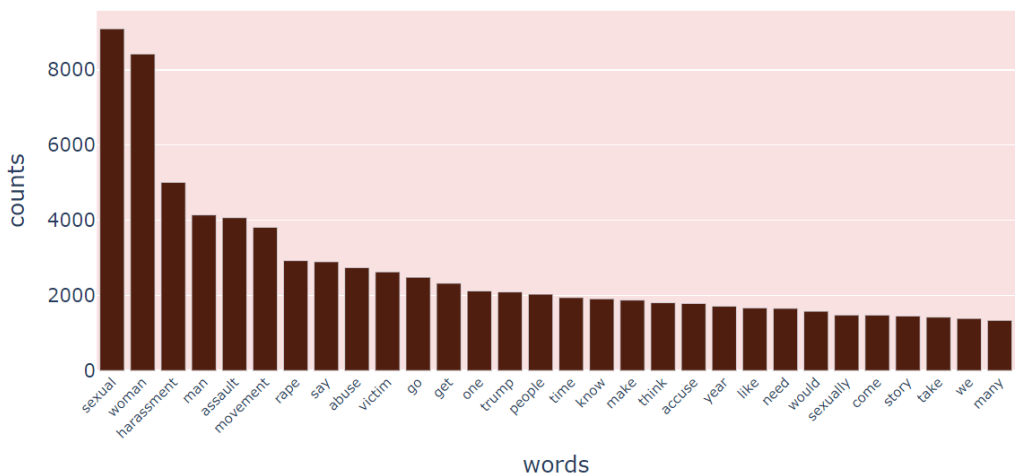
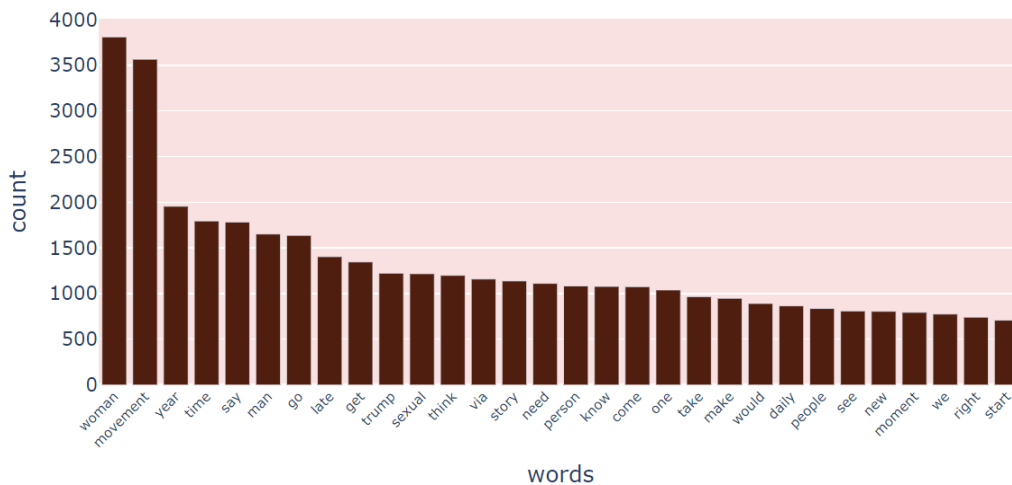Charts of top words sentiment-wise:

Top 30 Words in Positive Sentiment

Top 30 Words in Negative Sentiments

Top 30 Words in Neutral Sentiment

Topic Modeling

Through LDA (Latent Dirichlet Allocation) we have found 12 topics in the dataset. These topics show the variety and diversity in thoughts and opinions present in our dataset.

The Topics are as follows:
- Topic 0: This topic contains tweets about the political aspects of the movement such as congressmen, lawyers, senate members.
- Topic 1: Topic 1 contains tweets about empowering survivors, with ideas such as movies that highlight the plight of survivors, creating communities of strong people and discussing the role colleges and other institutions play in creating a "good" atmosphere
- Topic 2: Topic 2 deals with sexual harassment in workplaces such as offices as well as the mistreatment women face due to sexism and misogyny.
- Topic 3: Topic 3 talks about the impact of the #MeToo movement, calling the #MeToo movement "a great social change" as well as the trajectory of the movement.
- Topic 4: Topic 4 is about the time magazine's nominations for the 2017's person of the year where it had added and later named the "#MeToo movement" as person of the year.
- Topic 5: Topic 5 deals with prominent people in political parties and accusations on them with names like roy moore, Trump, lynne callahan being talked about multiple times.
- Topic 6: Topic 6 deals with people criticizing the movement as well as calling out people who are criticizing the movement.
- Topic 7: Topic 7 is about the demonstrations that were inspired by the #MeToo movement such as the clothesline project and changes brought about due to the movement.
- Topic 8: Topic 8 highlights survivor stories and talks about dr chris rocks and his books teaching children about appropriate touching.

- **Topic 9**: Topic 9 focuses on Tarana Burke the woman who created the movement in 2006 for empowering and uniting women of color to speak up about the abuse they faced
- **Topic 10**: Topic 10 deals with people talking about understanding sexual harassment & assault experiences and supporting women and girls and keeping them safe
- **Topic 11**: Topic 11 deals with addressing insufficient sex education and the efforts taken to bring justice to the victims of Sexual assault and harassment.

Conclusion

Through this in-depth exploratory data analysis we were able to gain insights into what happened in the movement and what kind of topics were dealt with. We understood how different public figures were involved and gauged the public's reactions to their involvement. The EDA has summarized the mass of tweets into charts and enabled us to pick apart and understand the small things that kept pushing the movement forward. We also were able to see the impact the #MeToo movement kept having on individuals. It gave them a platform, a voice which they used to get justice.

Overall the sentiment analysis provided an understanding of how public sentiment evolved since the beginning of the movement. In summary the analysis underscores the global impact and the also emphasizes the importance of addressing these issues.

The topics found during topic modeling cover a broad spectrum of discussions and each topic handles a distinct issue, even as there is some overlap and connections between them.

Link to GitHub Repository:

https://github.com/sakshii2004/MeToo-Movement-Analysis-Sentiment-Themes-and-Impact