International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES),(UGC APPROVED)
Impact Factor: 5.22 (SJIF-2017),e-ISSN:2455-2585
Research Symposium on
"Advancements in Engineering, Science, Management, and Technology"
Volume 5, Special Issue 04, April-2019.

Investigation of Tweets amid Me too Movement using Twitter Sentiment Analysis

Sherry Verma [1], Prakul Tomar[2], Aditi Jain [3], Manimala[4]

[1,2,3,4] *SET, Ansal University,*

Abstract*— *In recent years, expressing opinions and sharing ideas have become very easy, all with the help of social media. Social media platforms like Twitter, Facebook, etc. have also become a powerful tool to raise our voice against social injustices or problems faced by common people. In this research paper, a similar case has been considered. To analyse the voice raised by victims of sexual assault, sexual harassment, especially in workplace, is the main objective of this paper. Through MeToo movement twitter sentiment analysis we are finding the most affected places, most positive and negative word used, the part of day where incidents have occurred the most and the associated hashtags which are used with this movement for 6 days consecutively. For the analysis we have fetched tweets from twitter using python libraries, analysed them using Hive tool and for visualization matplotlib library is used.

Keywords*— *MeToo movement, sexual harassment social media, twitter sentiment analysis, and workplace

I. INTRODUCTION

Online social media has helped people to communicate, entertain, and share information over the internet. It has now focused majorly on user's interactions to see the weight of social statistics [1]. It has created a revolutionizing platform for people to communicate on highly stigmatized issues, which are there in our society [2]. Through social media, one can express his/her opinion on anything that has happened around the world or anything that person feels. Social media has a lot of influence on today's youth. In 2016, there were approximately 2.46 billion people using social media as stated by reports [6]. Twitter is one such platform where people communicate, use it as a business tool and share information. It is a microblogging service to send short messages online with brief content. The messages posted by users is called tweets [3]. It is also a channel to help users express their concerns about the disturbing things that happen with them or with others from time to time. Twitter was formed on 21st March 2016. Twitter has approximately 500 million tweets per day. On twitter, people make use of hashtags to draw attention, to organize or to promote. We can find various tweets using the same hashtag, as it allows exploring tweets regarding any event, any celebrity or just anything. Some of the hashtags starts becoming a trending topic when many people start using the same hashtag. These topics do not have any fixed duration of trend but rather depends on the use of that hashtag [3]. Sexual abuse is a major issue that has recently taken over social media by using the hashtag #MeToo with their shared posts [2]. MeToo movement begin in the year of 2006 but became recently popular in October 2016 when an American actress used this phrase and encouraged people to speak up. It is a movement which encourages the victims of sexual assault and sexual harassment usually conducted in workplace to stand up and help fight against the culprit [4]. Through this movement, many people from all over the world felt motivated and stood up against their culprit. Through ages, most of the women have been suppressed, denied their rights, become a target of abuse sexually or verbally, an object of entertainment and pleasure. However, through this movement many women have felt encouraged and motivated to share their story, to encourage others and to be with each other to stand up against this disgraceful thing. A recent study found out that more than 81 percent of women and approximately 43 percent of men in their lifetime have experienced some form of sexual harassment [5]. Through this research paper, we are visualizing some of the facts generated through our analysis on the MeToo movement using #MeToo hashtag on twitter. We have used Python libraries for fetching the tweets from twitter of 6 days. Analysis has been done using Hive tool from Big Data/Hadoop. This paper will be showing the top positive and negative words used in the timeline of 6 days starting from 20th October,2018 along with the most hashtags used by the people during this time. It will also show the environment like school, college or workplace abuse has occurred the most and the part of day when most of the incidents have occurred.

II. METHODOLOGY

A. Extracting Data from Twitter

Tweets of 6 days (duration: 20th October, 2018 to 25th October, 2018) has been fetched using tweepy library of python. We have fetched a total of 900mb size of tweets which consists of approx. 1, 07,350 tweets. Tweepy is an easy to use Python Library for accessing the Twitter API [7]. We have found the tweets using #MeToo. This data was downloaded in json format.

B. Preprocessing the data

The json data available to us was loaded into python notebook using pandas library. This helped in converting our unstructured data into structured one. In this dataset, we had 107836 rows and 37 columns. As most of the columns were unnecessary, we had removed 29 columns and was left with the following data for analysis. The data then obtained was saved in a csv file.

	entities	extended_tweet	filter_level	retweet_count	retweeted	retweeted_status	source	text
0	{'urls': [], 'hashtags': [{'indices': [81, 87]...	NaN	low	0	False	{'extended_tweet': {'entities': {'urls': [{'di...	<a href="http://twitter.com" rel="nofollow">Tw...	RT @brad_polumbo: Young men have a right to be...
1	{'urls': [], 'hashtags': [], 'user_mentions': ...	NaN	low	0	False	{'extended_tweet': {'extended_entities': {'med...	<a href="http://twitter.com/download/android" ...	RT @trehan_barkha: Share this video maximum ti...
2	{'urls': [{'display_url': 'twitter.com/i/web/s...	{'entities': {'urls': [], 'hashtags': [{'indic...	low	0	False	NaN	<a href="http://twitter.com/download/android" ...	@she_whispers Here I take on @bdutt who refuse...
3	{'urls': [{'display_url': 'twitter.com/i/web/s...	{'extended_entities': {'media': [{'display_url...	low	0	False	NaN	<a href="http://twitter.com" rel="nofollow">Tw...	@nikki_fortson @Jeff424V @jenjavajunky @Jennif...
4	{'urls': [], 'hashtags': [], 'user_mentions': ...	NaN	low	0	False	{'extended_tweet': {'entities': {'urls': [{'di...	<a href="https://mobile.twitter.com" rel="nofo...	RT @IJaising: I did NOT represent him on domes...
5	{'urls': [], 'hashtags': [], 'user_mentions': ...	NaN	low	0	False	{'extended_tweet': {'entities': {'urls': [], '...	<a href="http://twitter.com" rel="nofollow">Tw...	RT @Bit_2_close: None of his preys (hundreds o...
6	{'urls': [{'display_url': 'drive.google.com/op...	NaN	low	0	False	{'extended_entities': {'media': [{'display_url...	<a href="http://twitter.com" rel="nofollow">Tw...	RT @TheeDeepThroat: #CoryBooker Sexually Assau...

Fig 1. Data after Preprocessing

C. Word Count

The csv file was then fed as an input in Hive tool. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data and makes querying and analysing easy [8]. We had then split all the words present in text column using split function. As all the words were separated, we counted their total number of appearances in the entire tweet using group by clause and count function. The words were arranged in descending order to find the most occurred words. The sample output is as follows.

TABLE I. Count of most popular words

Word Count		
S.No	**Word**	**Count**
1	women	21652
2	Sexual	20536
3	night	18036

D. Analysis

In parallel to this, we have created our own dictionary assuming what most popular words we might find during our analysis. Addition to our own words, we have used words used on internet for past works [9] [10]. Now a comparison is being done between both the word counts i.e. the list obtained through word count and the count of words found from our dictionary. The most popular words have then been found and stored in a different file.

Similarly, for finding the most associated hashtags with MeToo movement, we have used word count using hashtag as keyword, then arranged them in descending order and saved it in a separate file.

For finding where and which part of the day has sexual abuse or assault happened more, we have created a word count of school, college and workplace for our analysis. After obtaining all these analyses, we have visualized them using matplotlib library of python. Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy [11].

III. **RESULT**

A. Frequency of incidents during a day

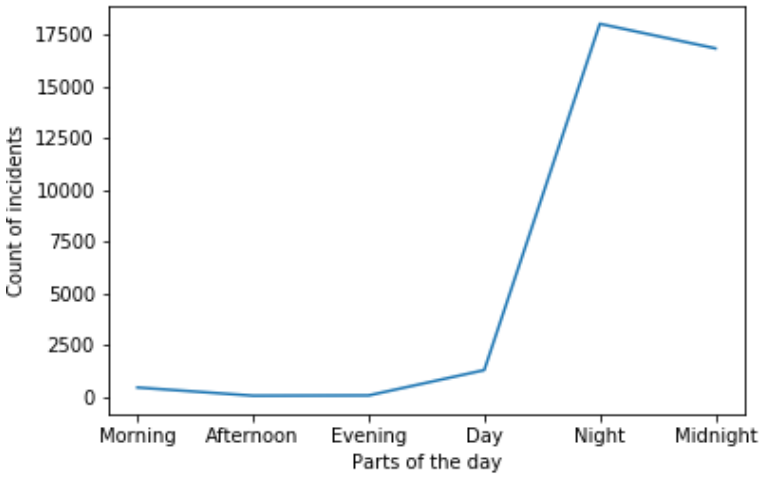

Fig 2. Frequency of incidents during different parts of the day

In this representation we are considering different parts of a day, the ratio between day and light is very alarming and is skewed towards night, as we can observer that the people have used words like morning, afternoon and evening very less. This indicates that more of harassment and assault has occurred during night and midnight.

B. Different places of incidents

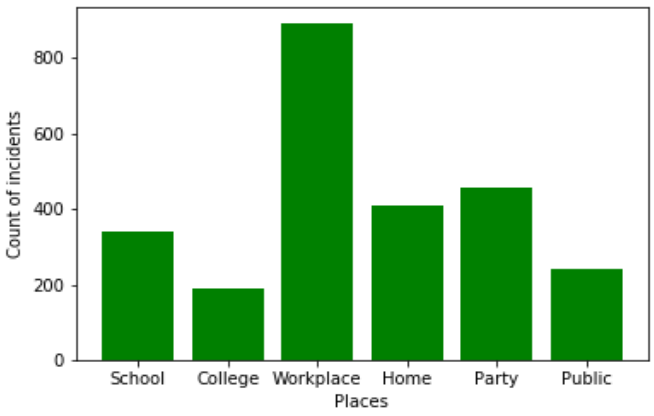

Fig 3. Frequenct of incidents at different places

Through this analysis, we have found that the most mentioned place in our data is Workplace; this indicates that most of the victims have faced the culprit daily. The second most mentioned place is Party and the most shocking thing that has been found is that home also has very high probability. When comparing school and college, the latter has least probability in our data.

C. Negative Words

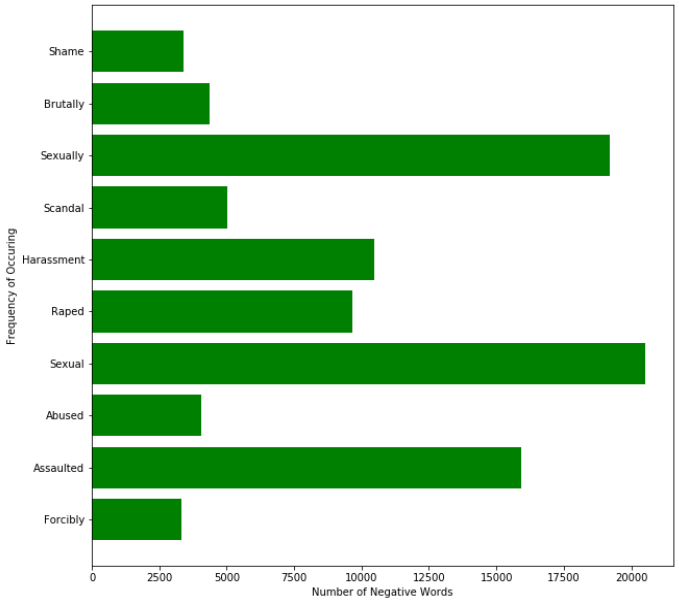

Fig 4. Popular negative words used by different users

Above graph show that sexual and sexually are most used negative words used and assaulted is the top 3rd word used by the people, which means that victims not only have to face sexually explicit statements, questions, jokes and abuse but is sexually assaulted and raped. Words like shame brutally and forcibly are also repeatedly.

D. Associative Hashtags used

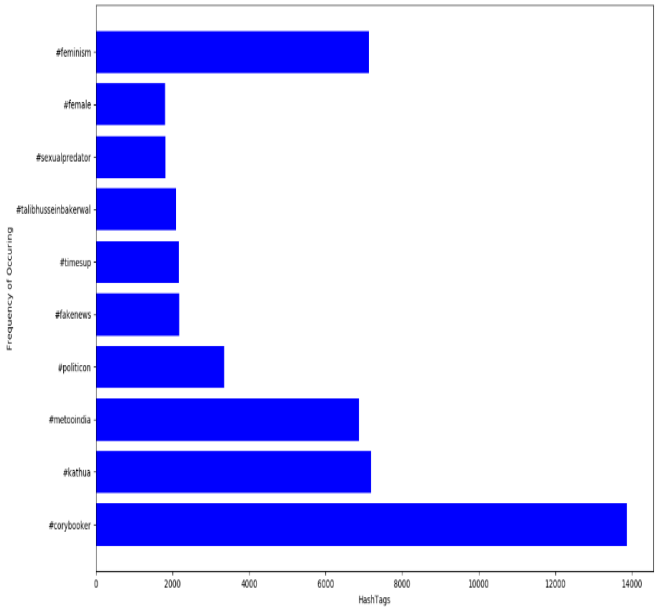

Fig 5. Most assosicated hashtags in Metoo tweets

As been observed the hashtag with maximum frequency is CoryBooker. Cory Booker is a US Senator who in early 1990's wrote in his book about groping a female friend's breast and she pushing his hand away. Due to his high-profile role, none could complain about him and no action could be taken [12]. The second most popular hashtag during these days was #kathua. Kathua is a village near Jammu Kashmir, here Asifa Bano, an 8-year-old-girl was abducted, gangraped, and was murdered coldblooded by some Hindu men [13]. The next popular hashtag is #metooindia. It appears as if authorities have to take this issue seriously, as there is so much nucleus that humungous people use a separate hashtag that is #metooinida. #feminism is also a very popular tag used by people as it indicates women wants equal rights and justice in society and some people used the hashtag #fakenews also which depicts that some people are misusing this movement in order to gain fame.

E. Emotive Words

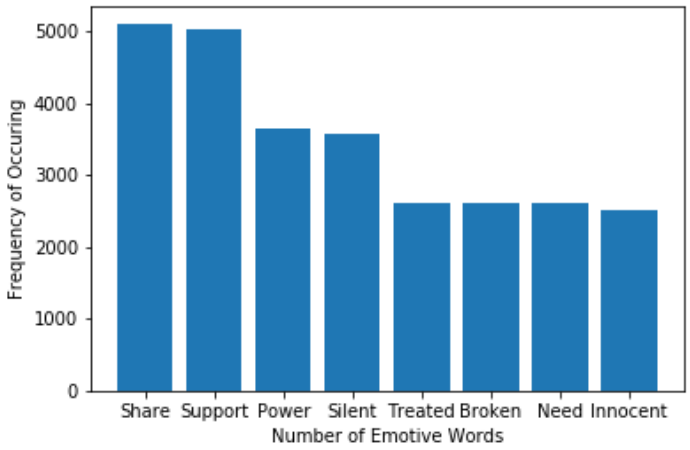

Fig 6. Emotive words used most frequently by the users

We have taken a count of some emotive words, which were used frequently by the users in their tweets. The two most popular words were share and support, which clearly indicates that everyone is trying to support and motivate each other to stand against their culprit. The users have also used words like need, innocent, and broken frequently.

IV. CONCLUSION

After performing this analysis, we would like to conclude that the number of cases is too high during the night-time and some action must be taken to avoid such thing. People have most suffered at their workplace, so office policies must be changed, and regulation should be done. Counselling sessions should be held to help people and make them understand what needs to be done. In addition, the top hashtags used shows a very disturbing outcome, which needs to be seen by authorities to let these stories be known to everyone. Comparing the use of negative and positive words we can see the frequency of negative words is very high which shouldn't be the case to make world a better place.

We can further extend our analysis by analysing tweets based on the geo-spatial location, so that we can figure out, which regions are most affected. Also, can study number of males and females affected. We can also find whether the tweets are real or fake depending on the user profile like counting their number of followers, friends and activity of user. By taking help of government, statistics can also be generated of how many crimes are reported in police department and how many just remain unheard.

REFERENCES

[1] Yun, Seokchan, Heungseok Do, and Hong-Gee Kim. "Analysis of user interactions in online social networks." Proceedings of the 19th International Conference on World Wide Web. 2010.

[2] Manikonda, Lydia, et al. "Twitter for Sparking a Movement, Reddit for Sharing the Moment:# metoo through the Lens of Social Media." arXiv preprint arXiv:1803.08022 (2018).

[3] Arroyo-Prado, V., et al. "Twitter Trending Topics and Accounts Analysis for a digital world understanding."

[4] https://en.wikipedia.org/wiki/Me_Too_movement

[5] https://www.npr.org/sections/thetwo-way/2018/02/21/587671849/a-new-survey-finds-eighty-percent-of-women-have-experienced-sexual-harassment

[6] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[7] http://www.tweepy.org/

[8] https://www.tutorialspoint.com/hive/

[9] https://public.tableau.com/profile/tamanna.hossain.kay#!/vizvizh/MeToo/MeToo

[10] https://data.world/marcmaxmeister/metoo-wordtree-corpus

[11] https://en.wikipedia.org/wiki/Matplotlib

[12] https://edition.cnn.com/2018/09/21/politics/cory-booker-brett-kavanaugh-sexual-assault-allegations/index.html

[13] Gonsalves, Roanna. "The crimson thread of male entitlement." Eureka Street 28.9 (2018): 45.