International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017),e-ISSN:2455-2585
International Conference on Smart Cities (ICSC-2019)
Volume-05, Special Issue_March_2019.

Insight into #MeToo using Sentiment Analysis

Aditi Jain
SET, Ansal University
Gurgaon, India

Prakul Tomar
SET, Ansal University
Gurgaon, India

Sherry Verma
SET, Ansal University
Gurgaon, India

Abstract* – Twitter is a very popular social media platform to address the problems faced by both common people as well as popular people. In recent years, me too movement has taken all over the twitter by the people belonging to all classes. This paper has been written to display the various important segments on this movement which might help the authorities to reduce the sexual crime rate. The purpose of this paper is to highlight the variations observed different segments of the Me-Too movement tweets over a month. This paper also shows the most popular tweets within the time period and the occupations having the maximum number of sufferers. For these predictions, Python libraries, Mongo DB and Map Reduce function has been used for analysis and visualization of the results.

***Keywords* – Map reduce, Me Too movement, Mongo DB, Sentiment analysis, Sexual abuse, Social media, Twitter**

I. INTRODUCTION

In recent years, it is said that social media has a great impact on communication and public discourse in the society [1]. Social media has been serving as a catchall phrase for web- based technologies conglomeration and for services like blogs, micro blogs (i.e., Wordpress, Twitter), text messaging, social sharing (e.g., YouTube, Flickr), discussion forums, collaborative editing and informative tools (e.g., wikis), and (e.g. Facebook, Instagram). These tools vary dramatically on their approaches and purposes, but they share emphasis on helping users to communicate, interact, and share content on the social environment [2]. Without any boundary or cost, it has helped most importantly common people to interact with people outside their scope. It has given encouragement for people to step out and speak and to influence people with their knowledge, creativity and words. Without social media platforms, the lives of many have changed positively as well as negatively. It has also created many job opportunities be it as creative writer, technical engineer, cyber specialist or creative writer. Our 65% of life now evolves and depend around these websites now.

Twitter is one such platform used for blogging and micro blogging. This micro blogging service twelve years old, Commands more than 335 million users as of October 2018 and is still growing. Users on Twitter can tweet about any topic with 280 characters as limit and follow others to see their content [3] [4]. It started a unique characteristic of putting a hashtag to create something trending. Twitter was ranked as twelfth most visited website by Alexa's web traffic analysis in March 2018 [5]. Twitter has played a big role in sentimental analysis and in prediction of results of various events. It allows us to fetch tweets from their server along with the additional information. For fetching tweets, various languages can be used such as python, Rlanguage and others.

In this paper, python libraries are used to fetch the data which is loaded into Mongo Db directly. Python is a programming language used for web development (server- side), scientific and numeric computation, software development, education, system scripting and others [6]. The next component we have used is Mongo Db. It is an open- source cross-platform document-oriented database language which uses JSON-like documents stored without a pre- defined schema [7]. Py Mongo python library has been used to fetch tweets from twitter. This library directly fetches the tweets and loads into Mongo db automatically. For calculating the count of words which is used for different purposes in this paper, we have used Map Reduce function of hive. It eases the word load, time and memory used by the system by 80% on an average. Matplotlib is another library of python which has been used in this paper for visualization of results.

This paper is presenting a twitter sentimental analysis on the Me-Too Movement. From years, women have been a subject of sexual abuse and violence. But now the time has come, that women have stood up and raised their voice to get justice, have an equal status in the society and motive and support others o stand against their culprits. Twitter sentiment analysis has been started using recently for general analyzation, prediction, classification and many others. Similarly, we have used twitter sentiment in analyzing different aspects of the currently trending me-too movement. Me too movement has taken a quick pace since October 2017, when an actress cam out in open and put allegations on a producer for misbehaving with her. Since then, a lot of

Women, mostly actresses have come out and told their stories. It had brought a movement throughout the world, supporting, encouraging and helping each other speak. Many unimaginative stories, incidents and people have come out through this movement. And, this brings to the aim of the paper to see what interesting results have come out in November, 2018 through this movement.

II. LITERATUREREVIEW

Sentiment analysis is a large developing area of Natural Language Processing with research from classification at document level (Turney,2002;Pang and Lee,2004), then at sentence level(hu and Liu,2004) to analyzing the phrase level(Wilson et al., 2005; Agarwal et al.,2009) [8] [10]. The first broad over view of sentiment analysis was shown in (Pang and Lee, 2008). Authors described the approaches and existing technologies used for opinion-oriented analyzing [9] in that paper. Blogs and micro blogs were not given much importance initially.

It is difficult and newer challenge to analyse twitter sentiment as users post reactions and opinions of everything. The very early and the recent results on sentiment analysis using tweets have been done by GO et al. (2009), Pak and Paroubek (2010) and by (Bermingham and Smeaton,2010) [10]. Classifiers such as Naive Bayes [12], Support Vector Machines [11] were used to build Models. Tweets ending with positive emoticons and negative emoticons have been used were used for analysis [10]. But only a little investigation has been done to check the importance of existing sentiment resources which was developed on non-micro blog data.

Recently researchers have now found a way to investigate ways of training the data automatically. Researchers like Pak and Paraoubak(2010), Barbosa and Feng (2010) have been relying on emoticons for defining their trained data. Usefulness of hashtags for creating training data has also been used, but they still limit experiments to sentiment/non- sentimental classes, rather than 3-way polarity classification [13], as generally been done[8].

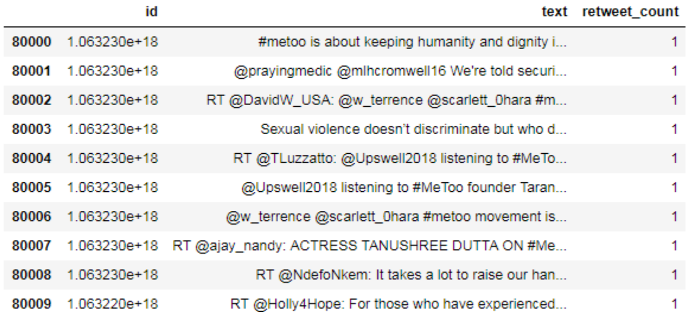
III. METHODOLOGY

A. ExtractingTweets

Tweets have been fetched using PyMongo[13] library of python. Approximately 1,07,350 tweets were fetched in the month of October and aaprox. 1,10,518 tweets in the Month of November. This has been done for comparison purposes. Tweets have been found using #meToo keyword. The data was directly loaded into Mongo automatically.

B. Preprocessing the data

After the fetching of tweets, they were then transferred to jupyter notebook for cleaning of data. It initially consisted of 110518 rows and 37 columns. For our analysis only 3 Columns were required, so the rest were cleaned out. The sample data is as follows.



	id	text	retweet_count
80000	1.063230e+18	#metoo is about keeping humanity and dignity i...	1
80001	1.063230e+18	@prayingmedic @mlhcromwell16 We're told securi...	1
80002	1.063230e+18	RT @DavidW_USA: @w_terrence @scarlett_0hara #m...	1
80003	1.063230e+18	Sexual violence doesn't discriminate but who d...	1
80004	1.063230e+18	RT @TLuzzatto: @Upswell2018 listening to #MeTo...	1
80005	1.063230e+18	@Upswell2018 listening to #MeToo founder Taran...	1
80006	1.063230e+18	@w_terrence @scarlett_0hara #metoo movement is...	1
80007	1.063230e+18	RT @ajay_nandy: ACTRESS TANUSHREE DUTTA ON #Me...	1
80008	1.063230e+18	RT @NdefoNkem: It takes a lot to raise our han...	1
80009	1.063220e+18	RT @Holly4Hope: For those who have experienced...	1

Fig 1. Sample Data set

C. Word Count

The csv file was then fed as an input in Hive tool for Map Reducing [14]. The words of text column had been separated using split () function. As all the words were separated, a total count was found using group by () clause and count () function. The words were then arranged in descending order to find the most popular words. The sample output is as follows.

Word Count		
Sl. No	**Word**	**Count**
1	women	9677
2	sexual	3778
3	against	3652

Table 1. Sample word count file

D. Maximum Retweet

For finding the most popular tweet over the twitter, we used retweet count column of the data to sort the in descending order. As the word limit available to us for downloading was just 140 characters, we looked up the tweet on the internet to provide better results.

E. Analysis

In parallel to this we 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 [15] [16]. 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, which part of the day and which occupation has been most affected, we searched our word count for different analysis. After obtaining all these analyses, we have visualized them using matplotlib library of python.

IV. RESULT

A. Occupations affected the most

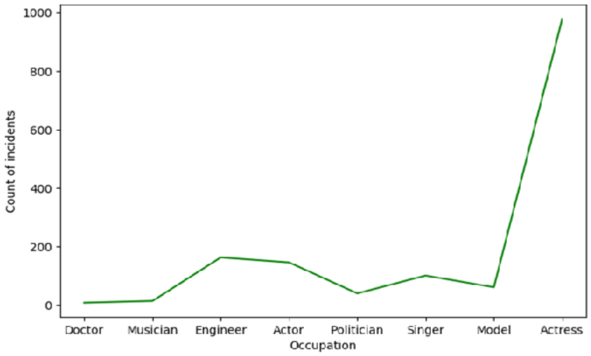


Fig 2. Count of incidents amongst popular occupations

Through this graph we are comparing different occupation that is affected most. We can see that actresses are the most affected and we have heard many issues are faced by actresses in film industry, the next most affected are engineers and actors. Singers and models also have a decent frequency.

B. Associative Hashtagsused

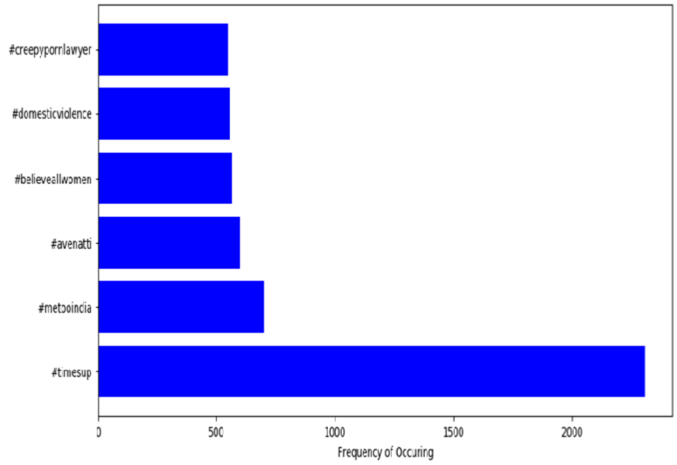


Fig 3. Most associated hashtags used

We have taken count of all hashtags and we have founded/found that #timesup is used by most people it is a movement against sexual harassment and was founded by Hollywood celebrities in response to Weinstein and #metoo. The second highest is the #metooindia, then 3rd highestis #avenatti he is an American attorney arrested for domestic violence.

C. Frequency of incidents during aday

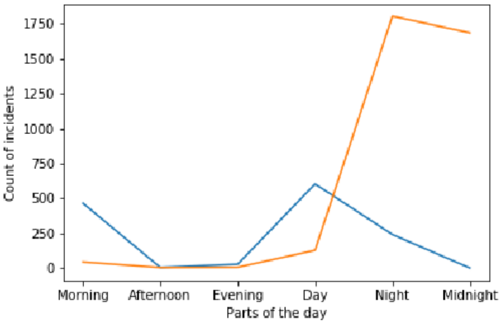


Fig 4. Comparison of number of incidents during different parts of the day

Here we are analyzing different parts of the day and we have taken 2-time intervals i.e October and November In the month of October we have seen the rate of incident is very high at night and midnight, but it is opposite in second dataset, the rate of incident is very high in day time. This is very alarming change that the crimes are done in day time more than night.

D. Different places of incidents

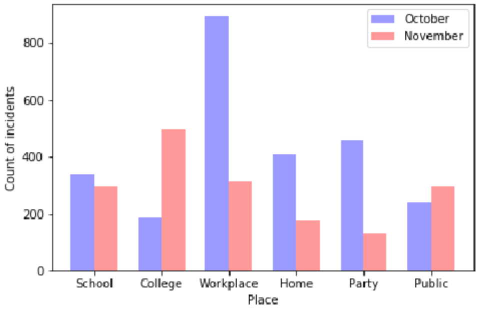


Fig 5. Number of incidents at different places

Here we have considered different places and we are comparing two different time intervals. In the October month the incidents are highest at workplace and in second dataset it is highest at college. School has observed the same number of cases. The ratio is decreased in case of home and party in second dataset.

F. Most popular Tweets

S. No.	Retweet Count	Username	Text
1	40310	Never On Brand	If a guy says he's nervous about #MeToo, just remind him that we come down pretty hard on murderers too, and ask him why that doesn't make him nervous. If he says, "Because I haven't murdered anyone," then you've learned something new about your friend.
2	25136	Kait Marieox	'Last night at the Trump rally in PA, a liberal protester touched me against my consent and said he'd throw me on the ground and rape me. Feminists around him, who say we're supposed to believe
3	18452	cnnbrk	When I raise my hand, I am aware of all the women who are still in silence." - Actress Viola Davis references the #MeToo movement during the Women's March in LosAngeles
4	17372	Alyssa_Milano	Don't let this performance fool you. If @SenatorCollins believed in #MeToo she would have opened the door when I was in her office to hear stories of constituent survivors.
5	10346	Rahul Gandhi	It's about time everyone learns to treat women with respect and dignity. I'm glad the space for those who don't, is closing. The truth needs to be told loud and clear in order to bringabout

Table 2. Most popular Tweets

VI. CONCLUSION AND FUTURE SCOPE

We have counted the top hashtags and analyzed different things like occupations, parts of the day, through our analysis we would like to conclude that number of cases too high in day time rather than night time and this is a serious issue as the culprits are not afraid of other people and authorities. College students have suffered the most, so the school authorities should investigate this matter seriously and awareness should be spread, and strict rules should be made. Also, the top hashtags used shows that a new movement like MeToo is emerging which is Times up and many users are using hashtags like #believeallwomen which depicts that some people still thinks that allegation made are fake. We can further extend our analysis by analyzing the data by location so that we can know where the risk is higher, and we can build a model which can predict the circumstances where the probability is high for a wrong event to occur. As we know people do no register complaints to authorities they just share on social media because of some reasons, we can provide help to those people who are afraid to talk by identifying them by their posts and activities.

REFERENCES

1. Stieglitz, Stefan, and Linh Dang-Xuan. "Social media and political communication: a social media analytics framework." *Social Network Analysis and Mining* 3.4 (2013): 1277-1291.
2. O'Keeffe, Gwenn Schurgin, and Kathleen Clarke-Pearson. "Clinical report—the impact of social media on children, adolescents, and families." *Pediatrics* (2011):peds-2011.
3. Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." *Proceedings of the 19th international conference on World Wide Web*. AcM,2010.
4. https://en.wikipedia.org/wiki/Twitter
5. https://www.alexa.com/siteinfo/twitter.com
6. https://www.w3schools.com/python/python_intro.asp
7. https://www.mongodb.com/what-is-mongodb
8. Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Icwsm*, *11*(538-541), 164.
9. Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp.1320-1326).
10. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*(pp. 30-38). Association for Computational Linguistics.
11. Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.
12. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*,*1*(12).
13. Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language*

processing (pp. 347-354). Association for Computational Linguistics.

14. Dittrich, J., & Quiané-Ruiz, J. A. (2012). Efficient big data processing in Hadoop MapReduce. *Proceedings of the VLDB Endowment*, *5*(12), 2014-2015.

15.https://public.tableau.com/profile/tamanna.hossain.kay#!/vizvizh/MeToo/MeToo

16.https://data.world/marcmaxmeister/metoo-wordtree-corpus