Analysis of Women Safety in Indian Cities Using Machine Learning on Tweets

Deepak Kumar[1], Shivani Aggarwal[2]

[1,2]*Amity University, Noida, India (Uttar Pradesh)*
[1]*aggarwalshivani32@gmail.com,* [2]*deepakgupta_du@rediffmail.com,*

Abstract: Women and girls have been experiencing a lot of violence and harassment in public places in various cities starting from stalking and leading to sexual harassment or sexual assault. This research paper basically focuses on the role of social media in promoting the safety of women in Indian cities with special reference to the role of social media websites and applications including Twitter platform Facebook and Instagram. This paper also focuses on how a sense of responsibility on part of Indian society can be developed the common Indian people so that we should focus on the safety of women surrounding them. Tweets on Twitter which usually contains images and text and also written messages and quotes which focus on the safety of women in Indian cities can be used to read a message amongst the Indian Youth Culture and educate people to take strict action and punish those who harass the women. Twitter and other Twitter handles which include hash tag messages that are widely spread across the whole globe sir as a platform for women to express their views about how they feel while we go out for work or travel in a public transport and what is the state of their mind when they are surrounded by unknown men and whether these women feel safe or not?

Keywords: Women, Safety, Sexual Harassment, Hash tag, Sentimental Analysis.

I. INTRODUCTION

There are certain types of harassment and Violence that are very aggressive including staring and passing comments and these unacceptable practices are usually seen as a normal part of the urban life. There have been several studies that have been conducted in cities across India and women report similar type of sexual harassment and passing off comments by other unknown people. The study that was conducted across most popular Metropolitan cities of India including Delhi, Mumbai and Pune, it was shown that 60 % of the women feel unsafe while going out to work or while travelling in public transport.

Women have the right to the city which means that they can go freely whenever they want whether it be too an Educational Institute, or any other place women want to go. But women feel that they are unsafe in places like malls, shopping malls on their way to their job location because of the several unknown Eyes body shaming and harassing these women point

Safety or lack of concrete consequences in the life of women is the main reason of harassment of girls. There are instances when the harassment of girls was done by their neighbours while they were on the way to school or there was a lack of safety that created a sense of fear in the minds of small girls who throughout their lifetime suffer due to that one instance that happened in their lives where they were forced to do something unacceptable or was sexually harassed by one of their own neighbor or any other unknown person.

Safest cities approach women safety from a perspective of women rights to the affect the city without fear of violence or sexual harassment. Rather than imposing restrictions on women that society usually imposes it is the duty of society to imprecise the need of protection of women and also recognizes that women and girls also have a right same as men have to be safe in the City.

Analysis of twitter texts collection also includes the name of people and name of women who stand up against sexual harassment and unethical behaviour of men in Indian cities which make them uncomfortable to walk freely. The data set that was obtained through Twitter about the status of women safety in Indian society was for the processed through machine learning algorithms for the purpose of smoothening the data by removing zero values and using Laplace and porter's theory is to developer method of analyzation of data and remove re-tweet and redundant data from the data set that is obtained so that a clear and original view of safety status of women in Indian society is obtained.

II. LITERATURE REVIEW

People often express their views freely on social media about what they feel about the Indian society and the politicians that claim that Indian cities are safe for women [1]. On social media websites people can freely Express their view point and women can share their experiences where they have faced sexual harassment or where we would have fight back against the sexual harassment that was imposed on them[2] . The tweets about safety of women and stories of standing up against sexual harassment further motivates other women data on the same social media website or application like Twitter. Other women share these messages and tweets which further motivates other 5 men or 10 women to stand up and raise a voice against people who have made Indian cities and unsafe

place for the women. In the recent years a large number of people have been attracted towards social media platforms like Facebook, Twitter and Instagram point and most of the people are using it to express their emotions and also their opinions about what they think about the Indian cities and Indian society. There are several method of sentiment that can be categorized like machine learning hybrid and lexicon-based learning. [5] Also there are another categorization Janta presented with categories of statistical, knowledge-based and age wise differentiation approaches. It is a common practice to extract the information from the data that is available on social networking through procedures of data extraction, data analysis and data interpretation methods. The accuracy of the Twitter analysis and prediction can be obtained by the use of behavioural analysis on the basis of social networks.

III. TWITTER ANALYSIS

As People communicate and share their opinion actively on social medias including Facebook and Twitter, Social network can be considered as a perfect platform to learn about people's opinion and sentiments regarding different events. There exists several opinion-oriented information gathering and analytics systems that aim to extract people's opinion regarding different topics. Since Twitter contains short texts, people tend to use different words and abbreviations. These phrases are difficult to extract their sentiment by current NLP systems easily. Therefore, many researchers have used deep learning and machine learning techniques to extract and mine the polarity of the phrases.

IV. IMPLEMENTATION OF SENTIMENTAL ANALYSIS OF TWEETS

In this technical paper, we will report the tweets picked up from Twitter API provided by Twitter itself. Due to the presence of Twitter API, there are many techniques available for sentimental analysis of data on Social media. In this project a set of available libraries has been used. The approach to extract sentiment from tweets is as follows:

1. Starting with downloading the sentimental dictionary

2. Then download the twitter testing data sets and add them as an input to the program.

3. Clean tweets by removing the stop words and noise like repetitive letters.

4. Tokenize each word and allot strength to the words in the dataset and feed it to the program.

5. For each word, compare it with positive sentiments and negative sentiments word dictionary and then increment positive count or negative count of the overall phrase.

6. Finally, based on the positive count & negative count, we can get result percentage about sentiment to decide the polarity which is categorized in Positive, Negative and Neutral.

Developers have done different sentimental analysis on Twitter for different purposes and a real-time twitter sentimental analysis of the trending events happening in the world, like elections, crimes, movies etc. Figure 1 shows the sentimental analysis algorithm at a higher level.

As it can be seen in the algorithm, we have different process to connect to the twitter API and fetch the tweets. Tweet cleaning or removal of stop words then classifying the tweets which means get the polarity of the tweet, and finally return the results shall be the later stage.

```
Algorithm 1 Extract Twitter sentiment
 1: procedure TWITTER-CONNECTION()
 2:     consumer − key =' xxxxxxxxx'
 3:     consumer − secret =' xxxxxxxx'
 4:     access − token =' xxxxxxxx'
 5:     access − token − secret =' xxxxxxxxxx'
 6:     self.auth = OAuthHandler(consumer − key, consumer −
 7:     self.auth.set − access − token(access − token, access − to
 8:     self.api = tweepy.API(self.auth)
 9: end procedure
10:
11: procedure TWEET-CLEANING(t)
12:     tweet = t.remove − Stop − words
13:     Return tweet
14: end procedure
15:
16: procedure TWEET-CLASSIFICATION(t)
17:     t = Tweet − Cleaning(t)
18:     tweet − polarity = t.sentiment.polarity
19:     tweet − polarity
20: end procedure
21:
22: procedure GET-TWEETS(q, count)
23:     fetched − tweets = self.api.search(q = query, count = co
24:     Return fetched − tweets
25: end procedure
```

Fig. 1. Sample code on a high level programming

A. Initial Setup

In this paper, we have used python to perform sentimental analysis. Some packages have been utilized including tweepy and textblob. We installed the required libraries by following commands:

1] pip install tweepy

2] pip install textblob

The second step is downloading the dictionary by running the following command:

- python -m textblob.download_corpora.

The textblob is a python library for natural language processing and it uses NLTK. Corpora is a structured set of texts/words which we need for analyzing the tweets.

B. Connecting to the Twitter API

To connect to the Twitter API and query latest tweets to store it in the database, we need to create an account on twitter and create an application. We were supposed to visit apps.twitter.com/app/new and generate the api keys required to feed the program. The application settings are shown in the figure2. Due to the security reasons the API keys are not shown.

Fig. 2. Twitter API application website

C. Result

Following shows the sample output of the program for the 'rape' as a query based on the last 300 tweets from Twitter.

- Positive tweets percentage: 16.39%

- Negative tweets percentage: 72.13%

- Neutral tweets percentage: 11.47%

Few Sample Tweets picked up from the database are:

<tweet>="@BinaNepram Sexual assaults on women #Manipur #NortheastIndia happening since armed conflict started 1960s~In a region where hundreds thousands armed forces operate under #AFSPA with impunity sanctioned by

parliament,NONE in #uniform accused of #rape punished in 50 years #MeTooIndia #MeToo" <tweet>="@Stateless1 Orphans, Rape Victims in refugee camp demand 'International Protection andJustice' #Refugees #WithRefugees #Rape #M yanmar #Bangladesh"

<tweet>="@So devastating to hear prominent figures named in

the #MeToomovement After some big names like #NanaPatekar and #AlokNath, now #KailashKher? scared of watching the news "

Blow is the final graph of the sentimental scores as per our implementation.

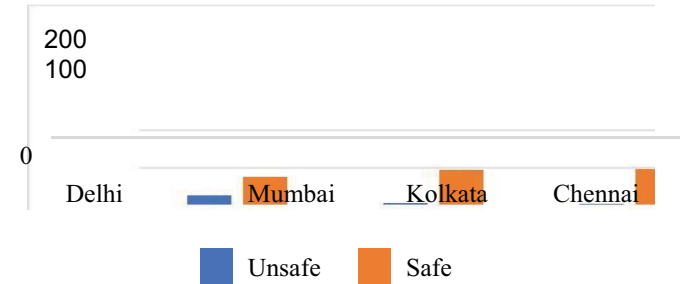City Safety chart for women based on tweets mined (in%)

TABLE 1: Saftey factor of population

Metro City	Cases tweeted about	Safety factor (tweet vs population)
Delhi	173,947	75%
Mumbai	42,940	93.6%
Kolkata	23,990	96.5
Chennai	13,442	98%

D. Final Report

If we run the program in various times we may get different results at every instance with a small variance, based on the tweets we fetch. We ran the program for three times and these results are the average of the consecutive outputs.

If the neutral tweets are significantly high, means that people have a lower interest in the topic and are not willing to haves a positive/negative side on it. This is also important to mention that depends on the data of the experiment we may get different results as people's opinion may change depending on the circumstances for example rape news it becomes the most trending news of the year in 2017. For some queries, the neutral tweets are more than 60% which clearly shows the limitation of the views. By above analysis that we have done, it can be clearly stated that Chennai is the safest city whereas Delhi is the unsafe city.

V. CONCLUSION

Throughout the research paper we have discussed about various machine learning algorithms that can help us to organize and analyze the huge amount of Twitter data obtained including millions of tweets and text messages shared every day. These machine learning algorithms are very effective and useful when it comes to analyzing of large amount of data including the SPC algorithm and linear algebraic Factor Model approaches which help to further categorize the data into meaningful groups. Support vector machines is yet another form of machine learning algorithm that is very popular in extracting Useful information from the Twitter and get an idea about the status of women safety in Indian cities.

REFERENCES

[1] Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009.

[2] Barbosa, Luciano, and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data." *Proceedings of the 23rd international conference on computational linguistics: posters*. Association for Computational Linguistics, 2010.

[3] Bermingham, Adam, and Alan F. Smeaton. "Classifying sentiment in microblogs: is brevity an advantage?." *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010.

[4] Gamon, Michael. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis." *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

[5] Kim, Soo-Min, and Eduard Hovy. "Determining the sentiment of opinions." *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

[6] Klein, Dan, and Christopher D. Manning. "Accurate unlexicalized parsing." *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003..

[7] Charniak, Eugene, and Mark Johnson. "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking." *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005.

[8] Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., & Badhani, P. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. *International Journal of Computer Applications*, *165*(9), 0975-8887.

[9] Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *2*(1), 178-183.

[10] Mamgain, N., Mehta, E., Mittal, A., & Bhatt, G. (2016, March). Sentiment analysis of top colleges in India using Twitter data. In *Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on* (pp. 525-530). IEEE.