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

Vijeth M
vijeth.5448@gmail.com
Reva University

Kirana G Y
kiranpatil9611@gmail.com
Reva University

Kiran Kumar B Hiremath
kiranh.2312@gmail.com
Reva University

Jayanth P
jayanthjeshu@gmail.com
Reva University

Prof. Thirumagal
thirumagal.e@reva.edu.in
Reva University

***Abstract*— From many decades women and girls have been facing a great deal of violence and harassment in many public areas of our country. From stalking to sexual harassment there would be a case regarding it daily. Here in our project, we focus on the function of social media in enhancing women's safety in Indian cities, here particularly we use Twitter platform. This project also focuses on how people of our country can establish a sense of responsibility in the safety of women in their environment. Twitter is one of the great platform that are widely used by the people and women to express their feelings about how they feel in their society for various activities such as travelling or surrounded by unknown peoples. Many people of our country, including media people and government institutions, are spreading the newest information and viewpoints on this issue. In this study, twitter data was extracted from the Twitter social media platform using the Python programming language and the Tweepy module is used to connect the twitter API and sentiment analysis was performed for those tweets collected using textblob, and then it was further analyzed.**

Keywords-component: Sentimental analysis, Twitter, TextBlob, Nltk, Tweepy.

I. INTRODUCTION

There are particular very aggressive kinds of harassment and violence, such as outstare and commenting, these intolerable actions are considered to be a normal scenes of a city. There have been various research that took place in many places of India, where people report sharing comments about the sexual harassment and other unidentified individuals. According to a survey conducted in India's most popular metropolitan areas such as Delhi, Mumbai, and Pune, 60% of women are at risk when going to work or using transport vehicles. In other words, you can walk around freely at any time, whether it's an educational institution or wherever a women feels to go. However, women sense unsafe because they feel embarrassed and harassed by multiple unfamiliar eyes, such as in malls or on their way to work.
The fundamental reason for girl harassment is a lack of safety or real consequences in the lives of women. There were many instances where women have been sexually harassed by their own neighbors when going to school, or where there has been a lack of safety that has created a sense of fear in the minds of small girls who have suffered throughout their lives as a result of that one incident where they were forced to do something unacceptable or were sexually harassed by one of their own neighbors or any other unknown person.

Safety of women in city is approached from the stance of women's rights, allowing them to influence the city without fear of violence or sexual harassment. Rather than imposing the restrictions that usually society imposes on girls, it's the duty of the society to understand the necessity for the safety of women and recognize women have the same right as the men has the safety, yet society has failed to recognize this.

Twitter is the most viable online entertainment stage to perform opinion examination. The kind of happy shared on every web-based entertainment stage differs with the highlights given by it. Facebook is a stage where the substance shared is now and again enormous in size and subsequently isn't reasonable for examination. Instagram is a site where the substance shared mostly centers around pictures and recordings as opposed to message. Consequently, when contrasted and the other web-based entertainment stages Twitter is the most appropriate systems administration stage since the substance shared on twitter is primarily text and emoticons. However, the substance from twitter is likewise not organized and not in the necessary structure for examination, it very well may be cleaned and handled effectively before use making it more appropriate than others.

Inspection of Twitter's text data also includes the names and the time of people who oppose sexual harassment and immoral behavior by men in various places of India. Repository on the status of women protection in Indian society, obtained via Twitter, are processed by machine learning algorithms, stripped of null values, analyze the data, and remove retweets and redundancy. We have developed a method to remove tags and smoothed the data. The data from the resulting dataset provides a clean and understandable picture of the safety status of women in Indian society.

We will primarily focus on the importance of the social media in boosting safety of women in India, where we primarily focus on the Twitter platform. We also ensure that following the end of this project, women and girls will have a better understanding of Indian cities.
With the use of social media, we can learn about the least safe places in our country and then educate people in that region to increase women's safety. As a result, we will make this world a better place to live for women.

II. LITERATURE REVIEW

[1] Here the author has introduced the sentimental analysis which is a method to recognize the emotion contained in the messages that are present in any virtual platform. It includes extracting keyword related tweets posted on Twitter. They have used the keyword #WorldEnvironmentDay to collect their dataset. They have used qualitative analysis software NVivo Pro 12 to arrange the sentiments as positive, negative, or neutral tweets.

[2] Here the author has done sentimental analysis which uses a python library in which testing, and training of data will not be required, so that the time consumption is less, hence we came across the new idea that is using a library called textblob. Before entering into textblob process, the collected tweets will be filtered that is all the unwanted tweets, emojis will be removed and in the textblob the tweets will be analyzed whether the tweet is positive or negative or neutral.

[3] The work in this paper is focusing on women and their safety them in various places. The method in this paper includes the use of social media Twitter which is a platform. It's used so that the user can get to know a clear idea of how safety works in the particular city. AI calculations were studied throughout the project. AI calculations are useful for sorting and investigating the tweets to classify.

[4] The vital part of this paper is to search the less secure states in Indian cities, hence they can use social media to educate people in the region to encourage women's safety. From which it will help to make this world a better place for women. Here the author focuses on efficiency analysis using three machine learning algorithms: Naive Bayes, Random Forest, and Support Vector Machines (SVMs). Here, they have got more accuracy by using the SVM algorithm. However, it takes time to train and test the model here.

[5] In this paperwork, they have taken Tweets collected from the twitter API as their database which they have used it as their input database, then preprocess those data sets (remove incomplete and noisy data) then they have applied feature extraction method finally used Naïve Bayes classification. This module has proposed an approach to detect sentiments on tweets using Naive Bayes classification
This classification method has low relevance for new information .

[6] Here the authors have used word embedding and lexicon-based application to classify the twitter data collected from the twitter API using tweepy. The collected dataset is taken as an input and then they are preprocessed by removing stop words, hash tags and other unwanted characters. Thereafter, the resultant text is tokenized & then processed with word embedding process to detect the location and lexicon-based method to find the emotion and sentiment of the tweets. But here Only Finite number of words in the lexicons are there.

III. SYSTEM REQUIREMENTS

The hardware requirements to run everything smoothly are as follows:

- Win 10 64bit.
- Min 4GB RAM/ Recommended 8GB.
- i5 8th Gen/ Ryzen 2nd Gen.
- Language: PYTHON.
- Tool: PyCharm.

IV. METHODOLOGY

In this project we have used python as a programming language and PyCharm as IDE to carry out the analysis. Before performing the sentiment analysis, we are required to install some required libraries. The required libraries can be installed using the following commands:

a. pip install tweepy

b. pip install matplotlib

c. pip install Django

d. pip install textblob

Tweepy is an open-source, easy-to-use Python package for accessing the Twitter API. Provides an interface for accessing API from Python applications.

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI

Django is a high-level Python-based free and open-source web framework that enables rapid development of secure and maintainable websites. that follows the model–template–views architectural pattern. It is maintained by the Django Software Foundation.

TextBlob is a package of python language for processing textual data. It has many in built functions to process common natural language processing (NLP) tasks such as Noun phrase extraction, sentiment analysis, classification, translation, etc.

The Sentiment property from the textblob package returns a named tuple in the form of Sentiment (polarity, subjectivity). The polarity score is a float in the range
[-1.0, 1.0]. Subjectivity is a float within the range of [0.0, 1.0], 0.0 is very objective and 1.0 is very subjective.

```
from textblob import TextBlob
sentence = 'The platform provides universal access to the world best education'

# Creating a textblob object and assigning the sentiment property
analysis = TextBlob(sentence).sentiment
print(analysis)
```

Once imported, we'll load in a text for analysis and instantiate a TextBlob object, as well as assigning the sentiment property to our own analysis.

Here the expected output of the analysis is:

```
Sentiment(polarity=0.5, subjectivity=0.26666666666666666)
```

Here the polarity is more than 0, so it is considered to be a positive comment.

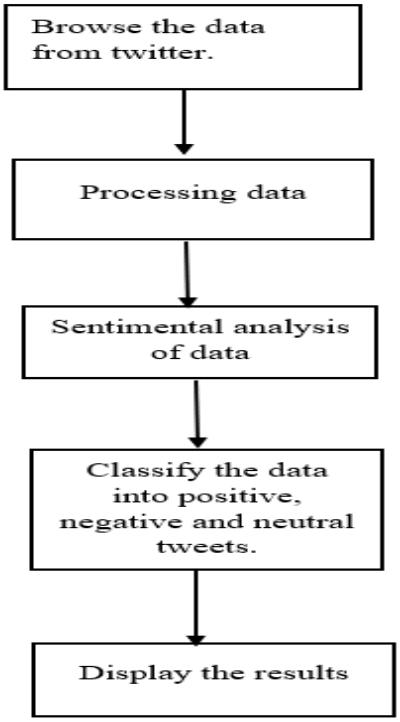
One of the great things about TextBlob is that it allows users to choose algorithms to implement high-level NLP tasks.

PatternAnalyzer is a standard classifier built on top of a pattern library.

NaiveBayesAnalyzer is an NLTK model trained in the movie review corpus.

Here we have used the default Pattern Analyzer for analyzing the sentiments of the tweets.

Analysis of the twitter data



The procedure for obtaining tweet estimations is divided into five stages:

1) ***Data Extractions :***
Here We basically connect out Twiiter API using the tweepy library and collect all the tweets from it. These are the data's which we are going to process further in our implementation part.

The Twitter API is a programming interface that provides a variety of tools for easy access to tweets. The Twitter API uses an HTTP to Get request to get a tweet.

By following these simple steps, you can easily connect to Twitter and retrieve data from Twitter.

1]. First, create / log in to your Twitter developer account.

2]. Click Create New App (enter all the required details on the Create App page).

3]. Now, the project will be created, once it is created, to get the consumer secret and consumer key click on "Keys and Access Tokens" tab. These consumer secrets and consumer keys will be used for authentication purposes.

2) ***Processing Data*** *:*
The data must be supplied to the classifier once it has been extracted from the twitter source as datasets. Here it cleans the dataset by removing unnecessary information such as stop words and emotes, ensuring that non-text-based substances are identified and removed before the study begins.

3) ***Sentimental analysis :***
Once the data is preprocessed , that is removing unwanted characters or unwanted tweets from the dataset than those resultant tweets are passed into our algorithm that is into our TextBlob python library, to analyse the sentiments expressed by the people on their tweets uploaded.

ALGORITHM:

```
1.  function Connect_with_Twitter()
2.      consumer_key = 'xxxxxxxxx'
3.      consumer_secret = 'xxxxxxxx'
4.      access_token = 'xxxxxxxx'
5.      access_token_secret = 'xxxxxxxx'
6.      self.auth = OAuthHandler(consumer_key, consumer_secret)
7.      self.auth.set_access_token(access_token, access_token_secret)
8.      self.api = tweepy.API(self.auth)
9.  end function
10.
11. function collect_tweets(tweet_count)
12.      collected_tweets = self.api.search(q=query,count=tweet_count)
13.      return collected_tweets
14. end function
15.
16. function clean_tweets(tweets)
17.     t= tweets.remove_stop_words
18.    return t
19. end function
20.
21. function classify_tweets(tweets)
22.      pre_processed_tweet = clean_tweets(tweets)
23.      tweet_polarity = pre_processed_tweet.sentiment.polarity
24.      Classify using tweet_polarity and Return value.
25. end function
```

4) ***Classifying the Data :***

Here it will classify each and every tweets present in the dataset whether it is positive, negative or neutral by using the PatternAnalyser algorithm which is pre built in python package called TextBlob. Than a separate count of these positive, negative and neutral tweets would be stored for further analysis.

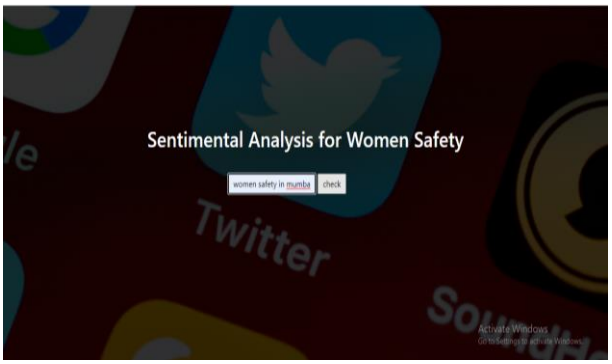
5) ***Displaying the results :***

Finally, the analysis and classification is displayed in an easily understandable manner.A graphical representation of all the classified tweets would be shown. Also the total count and the total percentile of the tweets would be displayed that is all the positive, negative and the neutral tweets present in the dataset collected.

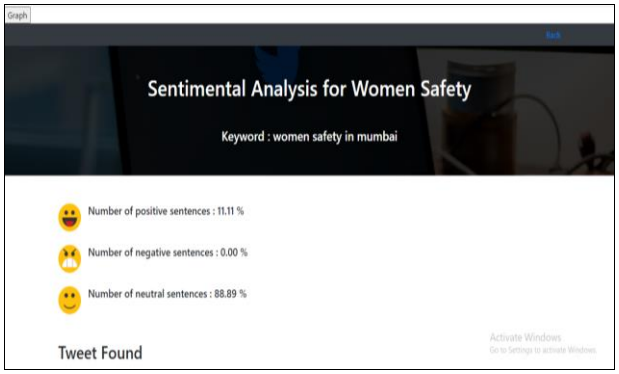
V. RESULTS AND DISCUSSIONS

With this project, we count the percentage of positive, negative, and neutral tweets by using the texblob python library for analyzing the tweets present in the dataset. Which will then be displayed graphically by using the matplotlib python library?

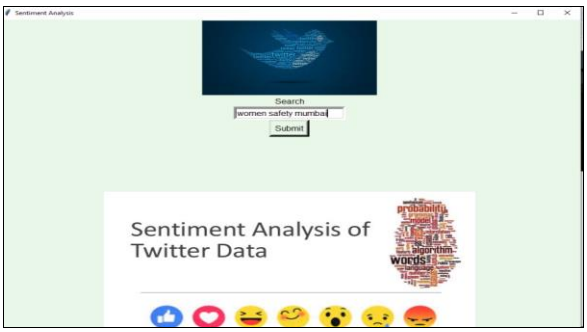
Once our project is executed the UI page that appears is as follows.



In the search bar we have to include the city name whose information is needed. Than after analysing everything in the backend the results would be displayed. The percentage of positive, negative, and neutral tweets for the keyword ‘women safety Mumbai’ will be as follows:

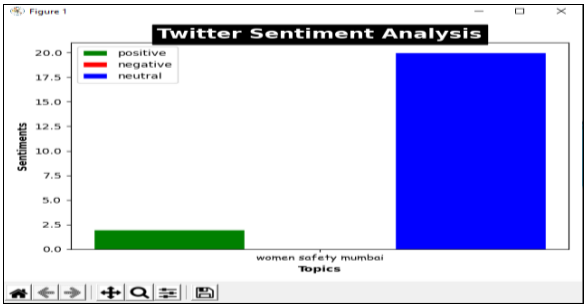


We Use the Emojis for representing Positive, Negative and Neutral tweets in our project. Also, the graphical representation will be shown for further analysis.



The bar graph indicates the percentage of positive tweets in green, percentage of negative tweets in red, and percentage of neutral tweets in blue.

The bar graph showing the positive, negative, and neutral percentages for Mumbai city is as follows:



Women and girls from outside India or different parts of India can use this website to check which city is safe for her to live in and will have a better understanding of Indian cities. Also creates awareness of which city is having more negativity of women harassment and the government can take necessary measures to overcome in that particular city and then educate people in that region to increase women's safety. As a result, we will make this world a better place to live for women.

VI. CONCLUSION

In this regard, the new technological environment, especially social media, Reflection of new technology development. Social media has become a new platform used by people, social movements, political parties, businesses, non-profits, or communities to share their opinion or concern on a particular topic.

We have discussed sentimental analysis which helps us to arrange and inspect the huge amount of Twitter dataset. The sentimental analysis using textblob is helpful and very effective when it comes to inspecting a huge amount of dataset.

Using this machine learning and social media platform, we can achieve sentimental analysis and bring more safety to women by spreading awareness. For future improvements, these machine learning technologies can be extended to apply to various web-based media platforms such as Facebook and Integral as our project work only considers Twitter.

VII. REFERENCES

[1] Reyes-Menendez, Ana, José Ramón Saura, and Cesar Alvarez-Alonso. "Understanding# WorldEnvironmentDay user opinions in Twitter: A topic-based sentiment analysis approach." International journal of environmental research and public health 15.11 (2018): 2537

[2] CHANDRA, VIKRAM, and RAMPUR SRINATH. "Analysis of Women Safety using Machine Learning on Tweets." (2020).

[3]D Swapna, Jampana Ashrita, Karpe Ashwini, Talasila Bindhu Bhargavi , "Analysis of Women Safety in Indian Cities Using Twitter Data." Journal Of Composition Theory (2021) ISSN : 0731-6755.

[4] Y Md, Riyazuddin & Sriram, G & Vaibhav, P & Vikranth, I. (2020). Utilization Of Support Vector Machine for Analyzing Women Safety in Indian States. International Journal of Grid and Distributed Computing. 2244-2251.

[5] Raparthi Shravya, Dr.P. Neelakantan, "Women Protection Analysis Based On Twitter Data Using ML"
European Journal of Molecular & Clinical Medicine, ISSN 2515-8260, 2020 .

[6] R. J. R. Raj, P. Das and P. Sahu, "Emotion Classification on Twitter Data Using Word Embedding and Lexicon Based Approach," 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT),2020,pp.150154,doi:10.1109/CSNT48778.2020.9115750.