Machine learning techniques to analyze tweets for insights into women's safety in Indian cities

Mullapudi Venkata Kiran Kumar[1], Mr.K. Subhash Chandra[2], Mallela Bhagya Sri[3]
Student[1, 3], Assistant.Professor[2]
Amritha Sai Institute of Science and Technology Paritala-521180
Autonomous NAAC with A Grade,Andhra Pradesh,India

ABSTRACT

Women and girls have been experiencing significant violence and harassment in public places across various cities, ranging from stalking to abuse, harassment, and assault. This research paper focuses on the role of social media in enhancing the safety of women in Indian cities, with particular emphasis on social media platforms such as Twitter, Facebook, and Instagram. Additionally, it explores how a sense of responsibility can be fostered within Indian society to prioritize the safety of women in their surroundings. Tweets on Twitter, which often include images, text, and quotes focusing on women's safety in Indian cities, can help convey important messages to the Indian youth and educate people to take strict actions against those who harass women. Twitter and other platforms with hashtag campaigns serve as a global platform for women to express their feelings about their safety while working, traveling on public transport, and being in the presence of unknown men, highlighting whether they feel secure or not.

Keywords: Women, Safety, Social Media, Tweets

I. INTRODUCTION

Certain types of harassment and violence, such as staring and passing comments, are very aggressive and are often considered a normal part of urban life. Several studies conducted in cities across India reveal that women frequently report similar types of sexual harassment and comments from strangers. One study conducted in major metropolitan cities, including Delhi, Mumbai, and Pune, found that 60% of women feel unsafe while going to work or traveling on public transport.

Women have the right to the city, meaning they should be able to move freely to educational institutions or any other place they wish to visit. However, many women feel unsafe in places like shopping malls or on their way to work due to body shaming and harassment from strangers. The lack of safety or concrete consequences for harassers is a primary reason for the harassment of women. Instances of harassment can occur even in their own neighborhoods, creating a lifelong sense of

fear in young girls who experience such traumatic events.

A "safe city" approach to women's safety considers women's rights to move about the city without fear of violence or sexual harassment. Instead of imposing restrictions on women, society must emphasize the need for their protection and recognize that women and girls have the same rights as men to feel safe in the city.

Analysis of Twitter texts includes identifying individuals who stand up against sexual harassment and unethical behavior in Indian cities. Data obtained from Twitter regarding women's safety in Indian society was processed using machine learning algorithms to smooth the data by removing zero values and redundant information, applying theories such as Laplace's and Porter's for data analysis. More than 500 million tweets are created daily on Twitter, making it the largest social network of its kind. Twitter's extensive user base encourages members to share their opinions on various topics, providing valuable insights for businesses, government agencies, and other organizations.

Using advanced machine learning techniques, this analysis seeks to categorize tweets based on content, sentiment, and location to provide a comprehensive understanding of the safety landscape in different Indian cities. By analyzing language, context, and user engagement, we aim to uncover both immediate and underlying factors that shape public perceptions and discussions related to women's safety.

II. LITERARURE SURVY

The literature on sentiment analysis of social media, particularly Twitter, offers valuable insights for understanding public perceptions of women's safety in Indian cities. Agarwal et al. [1] provide a method for contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams, which is crucial for interpreting nuanced expressions in tweets. Barbosa and Feng [2] address the challenges of sentiment detection on noisy and biased Twitter data, presenting a robust framework applicable to the complex nature of tweets related to women's safety. Bermingham and Smeaton [3] explore sentiment classification in the concise format of microblogs, highlighting the effectiveness of short texts in conveying sentiment. Gamon [4] emphasizes the importance of managing noisy data and large feature vectors in sentiment classification, which is essential for refining machine learning models to handle unpredictable Twitter data. Kim and Hovy [5] provide foundational

techniques for extracting sentiment from opinions, relevant for assessing public sentiment on women's safety. Klein and Manning [6] introduce accurate unlexicalized parsing techniques, enhancing syntactic analysis in sentiment classification, while Charniak and Johnson [7] propose a coarse-to-fine parsing approach that improves parsing efficiency and accuracy. Gupta et al. [8] conduct a comparative study on Twitter sentiment analysis using various machine learning algorithms, offering insights into selecting effective techniques for our analysis. Saha et al. [9] highlight the effectiveness of different machine learning approaches in sentiment analysis of Twitter data, and Mamgain et al. [10] demonstrate the applicability of sentiment analysis in evaluating public opinion on specific topics, such as educational institutions, which can be adapted to assess sentiments related to women's safety. Collectively, these studies provide a robust framework for analyzing Twitter data to uncover underlying factors shaping public discussions on women's safety, contributing to the development of strategies for improving safety measures and fostering a safer urban environment for women.

III. PROBLEM STATEMENT

EXISTING SYSTEM:

People often express their views freely on social media about their perceptions of Indian society and the politicians who claim that Indian cities are safe for women. Social media platforms provide a space where individuals can openly share their experiences, including instances of abuse and harassment, as well as stories of standing up against such mistreatment. Tweets about women's safety and accounts of resisting abuse and harassment serve to inspire other women on the same platform, such as Twitter. These shared messages and tweets create a ripple effect, encouraging more people to raise their voices against those who contribute to making Indian cities unsafe for women. In recent years, a significant number of people have turned to social media platforms like Facebook. It has become common practice to extract valuable information from social networking data through processes of data extraction, analysis, and interpretation. The accuracy of Twitter analysis and predictions can be enhanced by applying behavioral analysis based on social networks.

PROPOSED SYSTEM:

Women have the right to the city, meaning they should be able to move freely to educational institutions or any other place they wish to visit. However, many women feel unsafe in locations such as malls,

shopping centers, or on their way to work due to body shaming and harassment from strangers. The primary reason for this harassment is the lack of concrete consequences for perpetrators, leading to a pervasive sense of insecurity among women. Instances of harassment often occur even in familiar settings, such as by neighbors on the way to school, creating a lifelong sense of fear for young girls who endure such traumatic experiences. These experiences can profoundly affect their lives, as they suffer from the psychological impact of being forced into unacceptable situations or being abused by someone known to them or an unknown person.

Approaching women's safety from a perspective that emphasizes women's rights to navigate the city without fear of violence or harassment is crucial. Instead of imposing restrictions on women, which society often does, it is the duty of society to emphasize the importance of protecting women. Recognizing that women and girls have the same right as men to be safe in the city is essential for fostering a more secure and equitable urban environment.

Methodology:

Tweepy for Tweet Retrieval:

Utilizing the Tweepy package in Python, we fetched tweets from Twitter. Recognizing the intermittent availability of internet access, we preemptively downloaded tweets related to the #MeToo movement and women's safety, storing them in a designated dataset folder. This approach ensures that our application can access and analyze these tweets offline, enabling the detection of women's sentiments even without internet connectivity.

NLTK for Text Preprocessing:

To prepare the tweets for analysis, we employed the Natural Language Toolkit (NLTK). This tool facilitated the removal of special symbols and stop words from the tweets, ensuring that the text is refined and devoid of irrelevant elements. By cleansing the data in this manner, we aimed to enhance the accuracy of sentiment analysis by focusing solely on the meaningful content within the tweets.

TextBlob for Sentiment Analysis:

For sentiment analysis, we utilized the TextBlob library. Leveraging TextBlob's corpora package and dictionary, we quantified the polarity of each tweet, categorizing them as positive, negative, or neutral. Tweets with polarity values below 0 were classified as negative, those between 0 and 0.5 as neutral, and those exceeding 0.5 as positive. This systematic approach enabled us to discern and interpret the sentiments expressed in

tweets concerning women's safety effectively.

IV. RESULTS & DISCUSSION

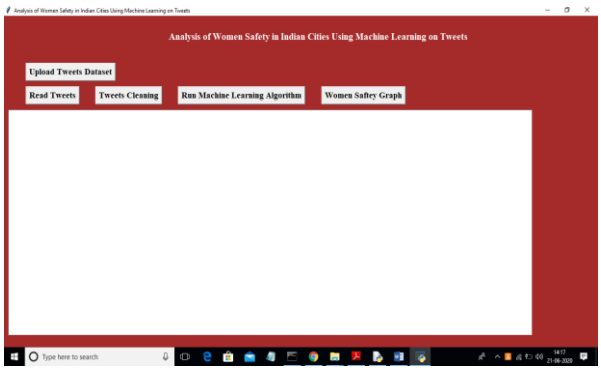

Fig-1. Upload Dataset

In above screen click on 'Upload Tweets Dataset' button and upload tweets

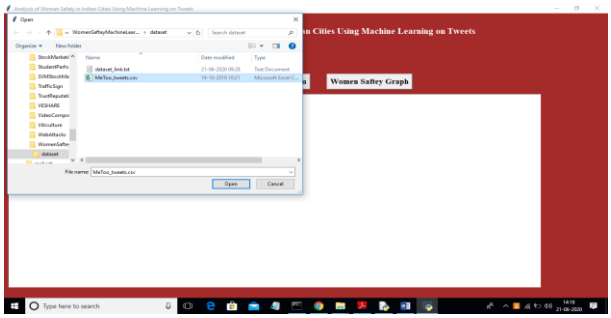

Fig-2. Select Dataset

In above screen uploading MeeToo_tweets.csv file and then click on 'Open' button to load dataset and to get below screen

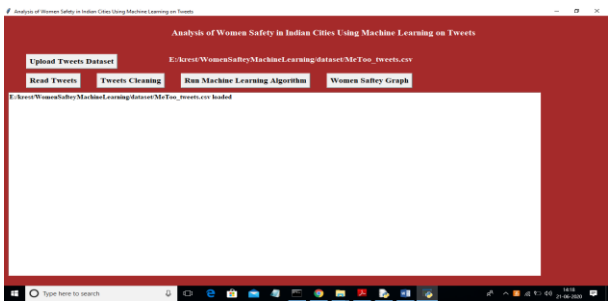

Fig-3. Read Tweets from Dataset

In above screen tweets dataset loaded and now click on 'Read Tweets' button to read tweets from dataset

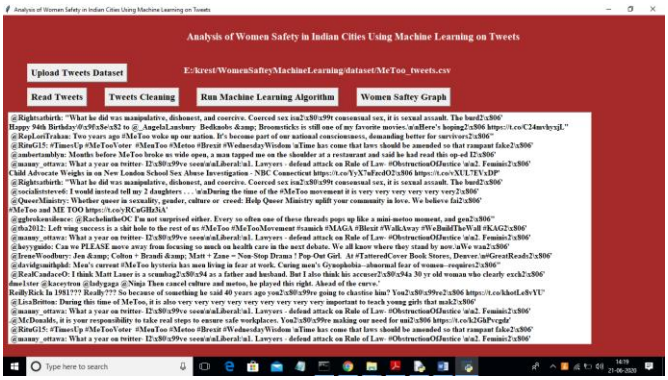

Fig-4. Tweets Loaded from Dataset

In above screen each line represents one tweet and you can scroll down above screen text area to view all tweets. In above screen we can see all tweets contains special symbols and stop words and to clean those tweets click on 'Tweets Cleaning' button

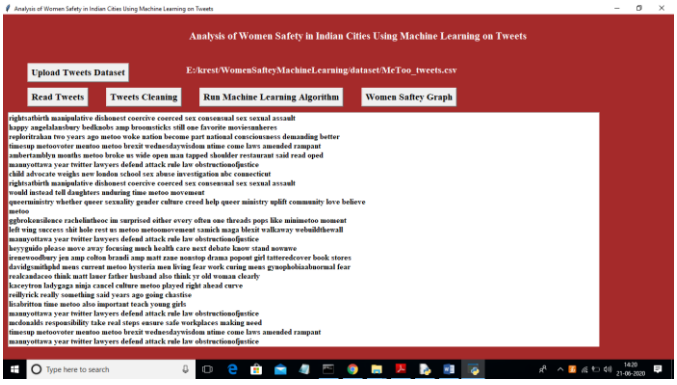

Fig-5. Dataset Cleaning

In above screen we can see all special symbols and stop words remove from tweets and only clean words are there and now click on 'Run Machine Learning

Algorithm' button to predict sentiments from tweets.

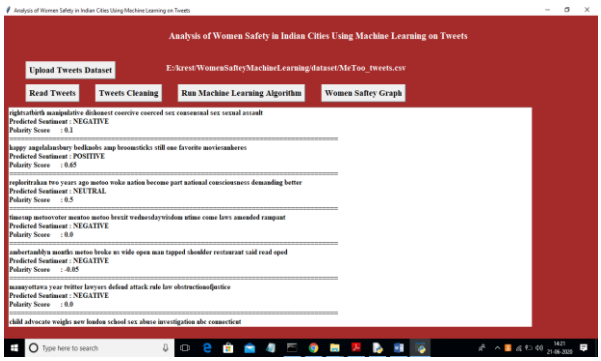

Fig-6. Tweets Prediction

In above screen each tweet having tweet text and then displaying tweets sentiments with polarity score. Scroll down above text area to see all tweets. Now click on 'Women Safety Graph' button to get known results and by seeing that result user can easily understand whether area is safe or not. If area is safe then more peoples will express either positive or neutral tweets and if not safe then more peoples will discuss negative tweets.

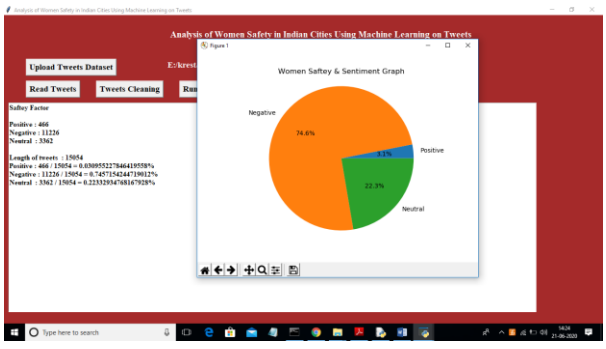

Fig-7. Women Safety Analysis Graph

In above screen 0.74 multiply by 100 will give 74% which means 74% peoples are talking negative and area is not safe and only 22 and 3% peoples are talking positive and neutral.

VI. CONCLUSION

Throughout our discussion, we have explored various machine learning algorithms that play a crucial role in organizing and analyzing the vast volume of Twitter data, encompassing millions of tweets and text messages shared daily. These algorithms prove to be highly effective and valuable in handling extensive datasets. Specifically, approaches such as the SPC algorithm and linear algebraic Factor Model contribute to further categorizing the data into meaningful groups, facilitating deeper insights. Additionally, support vector machines (SVM) emerge as another prominent machine learning algorithm renowned for extracting useful information from Twitter data. SVMs offer valuable insights into the status of women's safety in Indian cities by analyzing patterns and trends within the data.

VII. REFERENCE

[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] Saha yak, 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