# ANALYSIS OF WOMEN SAFETY IN INDIAN CITIES USING ML ON TWEETS

Rajasri M, Sreedevi B, Keshava Naidu N, Premnath S

Srinivasa Ramanujan Institute of Technology
Anantapur, India

**Abstract - Women and girls in Indian cities face significant challenges such as stalking, sexual harassment, and assault in public spaces. This research explores the role of social media platforms like Twitter, Facebook, and Instagram in promoting women's safety and raising awareness. Social media serves as a tool to disseminate messages, quotes, and stories, educating the youth and encouraging strict action against offenders. Women often use platforms like Twitter to express their feelings about safety while commuting or working, sharing their experiences through hashtags that reach a global audience.**

**To analyze these sentiments, this study employs machine learning algorithms, including Support Vector Machine (SVM), Neural Networks, Gradient Boosting, Random Forest, Decision Tree, Naive Bayes, and K-Nearest Neighbors (KNN). These models classify tweets into positive, negative, or neutral sentiments while assessing their performance using metrics like accuracy, precision, and recall.**

**This research highlights how technology and societal responsibility can promote women's safety. By leveraging machine learning and social media, it emphasizes the need for collective action to create safer environments for women**

**Keywords: Women's safety, social media, Twitter, machine learning, SVM, Neural Networks, Gradient Boosting, Random Forest, Decision Tree, Naive Bayes, KNN, sentiment analysis**

## I. INTRODUCTION

Twitter has become the most prominent microblogging platform, with over 100 million users and more than 500 million tweets sent daily. Its widespread reach makes it a powerful tool for individuals to share opinions, discuss topics, and engage in real-time conversations. As a result, Twitter serves as a valuable source of information for businesses, institutions, and organizations seeking to understand public sentiment and emerging trends.

With Twitter's 140-character limit, users often condense their messages using abbreviations, slang, emojis, and informal expressions. Many tweets also contain sarcasm, polysemy, and ambiguous language, making it difficult to interpret their exact meaning. This unstructured nature of Twitter content poses a challenge for traditional text analysis methods, requiring more advanced techniques to extract meaningful insights.

Sentiment analysis is a widely used approach to determine the emotions and opinions expressed in tweets. By categorizing tweets as positive, negative, or neutral, sentiment analysis helps in understanding public reactions to various events, products, policies, and social issues. It is an essential tool for businesses to assess customer feedback, for policymakers to gauge public sentiment, and for researchers to study social behavior.

To accurately analyze sentiments in Twitter data, machine learning techniques have been extensively explored. Various models and algorithms are employed to classify tweets based on sentiment, allowing for large-scale analysis of user opinions. These models improve decision-making processes in multiple sectors by providing real-time insights into trends and public perceptions.

One important area where sentiment analysis plays a crucial role is in addressing societal concerns, particularly issues related to women's safety. Many women use Twitter as a platform to share their experiences with harassment and violence, especially in urban areas. Studies indicate that women in cities such as Delhi, Pune, Chennai, and Mumbai often feel unsafe in public spaces, and Twitter provides a space to voice these concerns.

By analyzing tweets related to harassment and violence, sentiment analysis can help identify patterns and trends in public discourse. This information can be used to track locations where women feel unsafe, highlight recurring incidents, and understand the impact of social campaigns aimed at promoting women's safety. The ability to extract insights from these discussions makes sentiment analysis a valuable tool for addressing gender-based issues.

Moreover, tweets discussing harassment incidents often contain names of individuals involved, either as perpetrators or as people advocating for victims. By processing this data, researchers and policymakers can gain insights into the scale of such incidents and take appropriate action to improve public safety measures. Twitter thus acts as a digital record of real-world experiences, helping communities and authorities respond to pressing concerns.

Overall, sentiment analysis of Twitter data enables researchers to explore a wide range of topics, from political discourse to social justice issues. The vast amount of information available on the platform makes it an invaluable resource for studying public sentiment, shaping policies, and understanding societal challenges. As machine learning techniques continue to evolve, sentiment analysis will become an even more powerful tool for interpreting social media discussions and driving meaningful change.

## II. RELATED WORKS

In 2009, Apoorv Agarwal, Fadi Biadsy, and Kathleen R. McKeown presented "Contextual Phrase-Level Polarity Analysis Using Lexical Affect Scoring and Syntactic N-grams" at the 12th Conference of the European Chapter of the Association for Computational Linguistics. This research explored the use of syntactic n-grams and lexical affect scoring to analyze sentiment at the phrase level, improving sentiment classification accuracy.

In 2010, Luciano Barbosa and Junlan Feng published "Robust Sentiment Detection on Twitter from Biased and Noisy Data" at the 23rd International Conference on Computational Linguistics. Their study focused on improving sentiment detection on Twitter by addressing biased and noisy data using machine learning techniques.

Adam Bermingham and Alan F. Smeaton, in 2010, investigated the impact of text brevity on sentiment classification in microblogs. Their work, "Classifying Sentiment in Microblogs: Is Brevity an Advantage?," was presented at the 19th ACM International Conference on Information and Knowledge Management. It examined whether shorter texts, such as tweets, affect the accuracy of sentiment analysis.

Michael Gamon, in 2004, explored sentiment classification in customer feedback data in his research "Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis." Presented at the 20th International Conference on Computational Linguistics, this study analyzed the challenges posed by noisy data and large feature vectors in sentiment classification.

In 2004, Soo-Min Kim and Eduard Hovy introduced "Determining the Sentiment of Opinions" at the 20th International Conference on Computational Linguistics. Their research investigated techniques for extracting sentiment from opinionated text using computational linguistics.

Dan Klein and Christopher D. Manning presented "Accurate Unlexicalized Parsing" at the 41st Annual Meeting of the Association for Computational Linguistics in 2003. This study focused on improving parsing accuracy using unlexicalized probabilistic models, which play a significant role in text processing for sentiment analysis.

Eugene Charniak and Mark Johnson, in 2005, introduced "Coarse-to-Fine N-Best Parsing and MaxEnt Discriminative Reranking" at the 43rd Annual Meeting of the Association for Computational Linguistics. This research emphasized parsing efficiency improvements, contributing to better natural language processing techniques.

In 2017, B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, and P. Badhani conducted a study titled "Study of Twitter Sentiment Analysis Using Machine Learning Algorithms on Python," published in the International Journal of Computer Applications. Their work analyzed the effectiveness of machine learning algorithms in sentiment classification.

In 2015, V. Sahayak, V. Shete, and A. Pathan explored sentiment analysis techniques in their study "Sentiment Analysis on Twitter Data," published in the International Journal of Innovative Research in Advanced Engineering (IJIRAE). Their research examined various methods used to extract sentiment from Twitter content.

N. Mamgain, E. Mehta, A. Mittal, and G. Bhatt presented "Sentiment Analysis of Top Colleges in India Using Twitter Data" at the International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT) in 2016. Their study analyzed Twitter sentiment to evaluate public perceptions of higher education institutions in India.

## III. EXISTING SYSTEM

The previous systems used for crime detection, particularly in addressing crimes against women, were highly reliant on manual processes. Police and authorities had to manually search through physical records to gather necessary information, which was a labor-intensive and time-consuming task. This lack of automation led to significant delays in identifying criminals or responding to urgent situations. The manual system made it difficult to track incidents in real-time, and critical information could easily be overlooked, resulting in slow response times and ineffective intervention.

Additionally, the old system lacked technological integration, which further complicated matters. Without the use of advanced tools and data analytics, law enforcement officials had no quick means of accessing or analyzing large volumes of data. As a result, authorities were often unaware of patterns of criminal behavior, and important connections between incidents could go unnoticed. The absence of automated alerts or real-time monitoring made it difficult to provide timely assistance to victims, particularly women facing harassment or violence in urban environments.

One of the key issues with the previous system was the challenge of identifying suspects or tracking down criminals. Data was not centralized, and there was no unified database that could provide instant access to critical information. Investigators had to rely on manual searches across different departments, which was not only inefficient but also prone to human error. This inefficiency delayed investigations and often meant that crucial details were missed, hindering the effectiveness of the justice system in protecting women.

In summary, the existing system had several significant limitations: it was slow, inefficient, and prone to errors. The manual process of searching for records, combined with the lack of technological tools to aid investigations, meant that crimes were not addressed quickly enough, leading to a higher risk for victims. To overcome these challenges, there was a pressing need for more advanced systems that could automate processes, provide real-time data, and enhance the efficiency and accuracy of law enforcement in preventing and solving crimes, especially those related to women’s safety.

## IV. PROPOSED SYSTEM

In light of the rapid changes in society and the increasing awareness of crimes, particularly those affecting women, there is a growing need for an efficient and modern approach to crime detection. The proposed system aims to address this need by automating crime record management and utilizing advanced machine learning (ML) algorithms. With this system in place, authorities can swiftly retrieve and analyze criminal records, making it easier to identify criminals, track their history, and take appropriate action without the delays of manual record-keeping.

The system utilizes several powerful machine learning algorithms, including AdaBoosting, CatBoosting, Support Vector Machines (SVM), and Naïve Bayes. These algorithms are designed to train the dataset and predict outcomes based on historical information. By processing large volumes of data, the system can quickly identify patterns of criminal behavior, helping authorities spot potential offenders faster and more accurately. In addition, it allows for seamless access to past records, making it easier to track a suspect's criminal history.

One of the key benefits of this approach is the significant reduction in time required to process and retrieve criminal data. Traditional methods of record-keeping and search are time-consuming and often ineffective, but with ML algorithms, the system can process data in real-time, enabling faster decision-making. This improved response time can lead to quicker intervention and more efficient handling of criminal cases, reducing the opportunity for crimes to escalate.

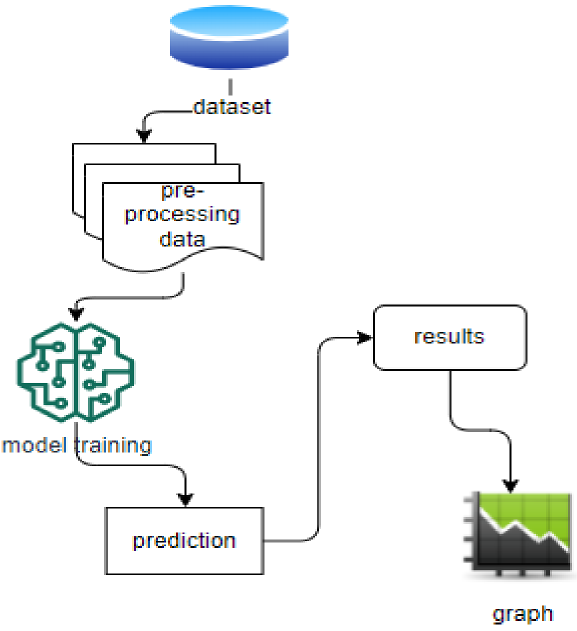
Moreover, the system's reliance on advanced algorithms ensures that the results produced are highly accurate. Unlike manual searches that may be prone to human error, the machine learning algorithms continuously improve over time, leading to more precise predictions and classifications. This not only enhances the effectiveness of crime detection but also ensures that the data-driven insights guide decisions with minimal bias or error.

Another advantage of the proposed system is its overall efficiency. By automating processes such as record-keeping, search, and analysis, the workload of authorities is significantly reduced. This allows law enforcement to focus on more pressing tasks while the system handles the heavy lifting of data processing. This increased efficiency can lead to a more effective allocation of resources, improving the overall effectiveness of the justice system.

In conclusion, the proposed system leverages cutting-edge machine learning technologies to enhance the speed, accuracy, and efficiency of crime detection and prevention. By reducing the time needed for data retrieval, improving the accuracy of predictions, and streamlining processes, this system offers a comprehensive solution to addressing crime in a modern, data-driven manner. Through these advancements, the system not only benefits law enforcement agencies but also contributes to creating a safer environment for society, particularly women.

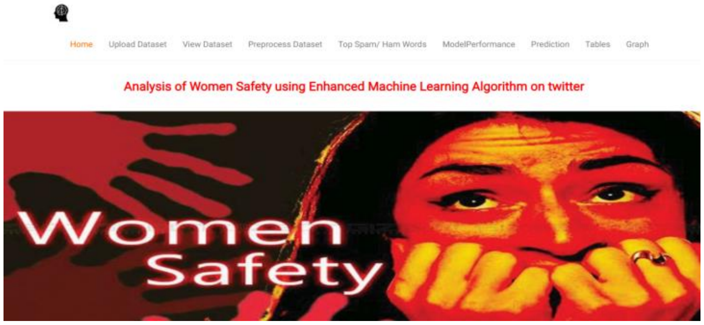
Additionally, the system's ability to continuously learn and adapt ensures that it stays relevant as crime patterns evolve, further improving its long-term effectiveness. With this adaptability, the system can become a powerful tool for proactive crime prevention and a crucial resource for both law enforcement and community safety.

## V. ARCHITECTURE



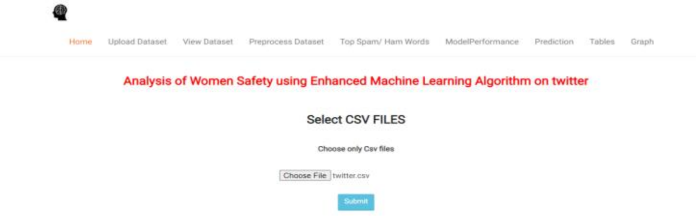
The process starts with uploading a dataset, followed by pre-processing to clean and structure the data. Next, model training is performed using machine learning algorithms. The trained model generates predictions, which are analyzed to produce results. Finally, the results are visualized using a graph, providing insights into data trends.

## VII. RESULT AND DISCUSSION



Analysis of Women Safety using Enhanced Machine Learning Algorithm on twitter

The homepage serves as a central hub for uploading datasets, preprocessing data, analyzing trends, and making predictions on women's safety-related tweets using machine learning.



Analysis of Women Safety using Enhanced Machine Learning Algorithm on twitter

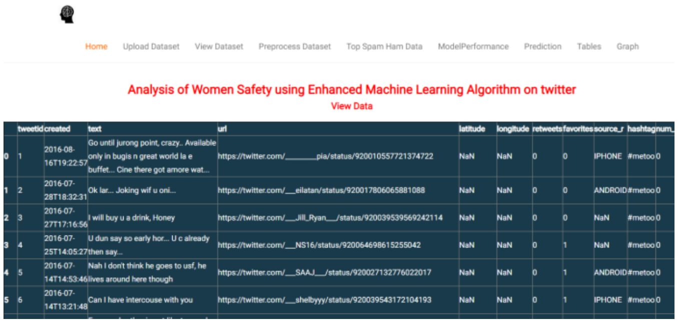
Select CSV FILES

Choose only Csv files

twitter.csv

The "Upload Dataset" page allows users to upload CSV files for analysis related to women's safety on Twitter.

The "View Dataset" page displays the uploaded dataset in a tabular format for review and analysis.

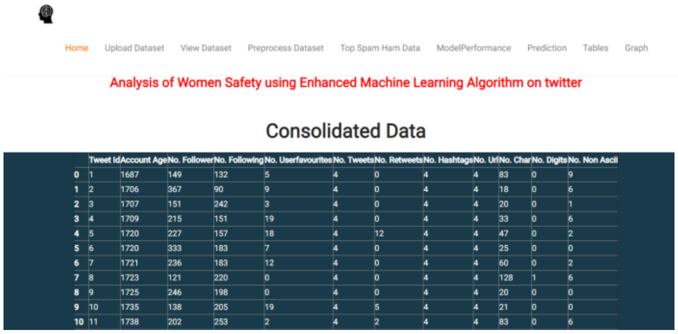


Analysis of Women Safety using Enhanced Machine Learning Algorithm on twitter

View Data

| | tweetid | created | text | url | latitude | longitude | retweets | favorites | source_r | hashtag | num_ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2016-08-16T19:22:57 | Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat... | https://twitter.com/________pia/status/920010557721374722 | NaN | NaN | 0 | 0 | IPHONE | #metoo | 0 |
| 1 | 2 | 2016-07-28T18:32:31 | Ok lar... Joking wif u oni... | https://twitter.com/___eilatan/status/920017806065881088 | NaN | NaN | 0 | 0 | ANDROID | #metoo | 0 |
| 2 | 3 | 2016-07-27T17:16:56 | I will buy u a drink, Honey | https://twitter.com/___Jill_Ryan___/status/920039539569242114 | NaN | NaN | 0 | 0 | NaN | #metoo | 0 |
| 3 | 4 | 2016-07-25T14:05:27 | U dun say so early hor... U c already then say... | https://twitter.com/___NS16/status/920064698615255042 | NaN | NaN | 0 | 1 | NaN | #metoo | 0 |
| 4 | 5 | 2016-07-14T14:53:46 | Nah I don't think he goes to usf, he lives around here though | https://twitter.com/___SAAJ___/status/920027132776022017 | NaN | NaN | 0 | 1 | ANDROID | #metoo | 0 |
| 5 | 6 | 2016-07-14T13:21:48 | Can I have intercouse with you | https://twitter.com/___shelbyyy/status/920039543172104193 | NaN | NaN | 0 | 1 | IPHONE | #metoo | 0 |

The "Preprocess Dataset" page displays the cleaned and consolidated dataset with various Twitter metrics for analysis.

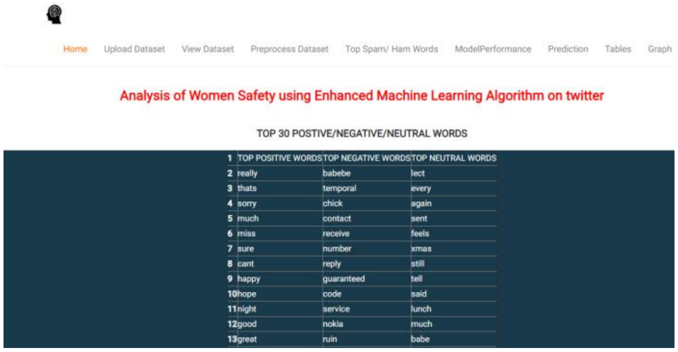


Analysis of Women Safety using Enhanced Machine Learning Algorithm on twitter

Consolidated Data

| | Tweet Id | Account Age | No. Follower | No. Following | No. Userfavourites | No. Tweets | No. Retweets | No. Hashtags | No. Url | No. Char | No. Digits | No. Non Ascii |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1687 | 149 | 132 | 5 | 4 | 0 | 4 | 4 | 83 | 0 | 9 |
| 1 | 2 | 1706 | 367 | 90 | 9 | 4 | 0 | 4 | 4 | 18 | 0 | 6 |
| 2 | 3 | 1707 | 151 | 242 | 3 | 4 | 0 | 4 | 4 | 20 | 0 | 1 |
| 3 | 4 | 1709 | 215 | 151 | 19 | 4 | 0 | 4 | 4 | 33 | 0 | 6 |
| 4 | 5 | 1720 | 227 | 157 | 18 | 4 | 12 | 4 | 4 | 47 | 0 | 2 |
| 5 | 6 | 1720 | 333 | 183 | 7 | 4 | 0 | 4 | 4 | 25 | 0 | 0 |
| 6 | 7 | 1721 | 236 | 183 | 12 | 4 | 0 | 4 | 4 | 60 | 0 | 2 |
| 7 | 8 | 1723 | 121 | 220 | 0 | 4 | 0 | 4 | 4 | 128 | 1 | 6 |
| 8 | 9 | 1725 | 246 | 198 | 0 | 4 | 0 | 4 | 4 | 20 | 0 | 0 |
| 9 | 10 | 1735 | 138 | 205 | 19 | 4 | 5 | 4 | 4 | 21 | 0 | 0 |
| 10 | 11 | 1738 | 202 | 253 | 2 | 4 | 2 | 4 | 4 | 83 | 0 | 6 |

The "Top Spam/Ham Words" page displays the top 30 positive, negative, and neutral words identified from the dataset for spam and sentiment analysis.



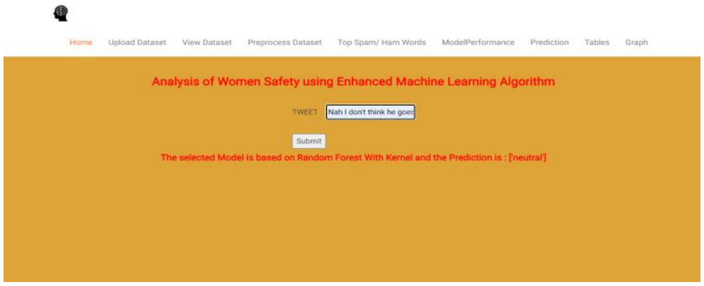
Analysis of Women Safety using Enhanced Machine Learning Algorithm on twitter

TOP 30 POSTIVE/NEGATIVE/NEUTRAL WORDS

| 1 | TOP POSITIVE WORDS | TOP NEGATIVE WORDS | TOP NEUTRAL WORDS |
|---|---|---|---|
| 2 | really | babebe | lect |
| 3 | thats | temporal | every |
| 4 | sorry | chick | again |
| 5 | much | contact | sent |
| 6 | miss | receive | feels |
| 7 | sure | number | xmas |
| 8 | cant | reply | still |
| 9 | happy | guaranteed | tell |
| 10 | hope | code | said |
| 11 | night | service | lunch |
| 12 | good | nokia | much |
| 13 | great | ruin | babe |

The "Model Performance" page displays the performance metrics (accuracy, precision, and recall) of the selected machine learning model, whether it's based on SVM, Neural Network, Gradient Boost, Random Forest, Decision Tree, Naive Bayes, or K-Nearest Neighbors (KNN). Each model's metrics are showcased to provide insights into its effectiveness for the given task.

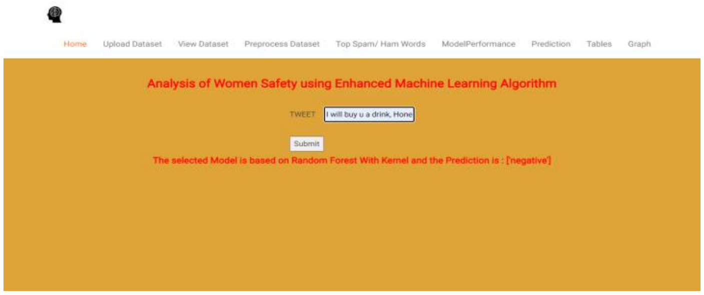
The "Prediction" page allows users to input a tweet and analyze it using the selected machine learning model, providing a sentiment classification such as neutral, negative, or positive, depending on the model used (e.g., Random Forest). The sentiment prediction is based on the analysis of the tweet's content.



Analysis of Women Safety using Enhanced Machine Learning Algorithm

TWEET: Nah I don't think he goes

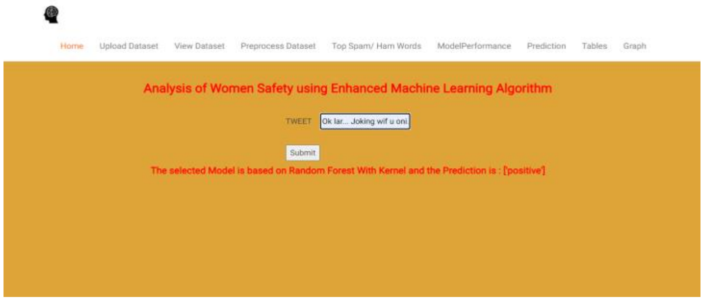
The selected Model is based on Random Forest With Kernel and the Prediction is : ['neutral']



Analysis of Women Safety using Enhanced Machine Learning Algorithm

TWEET: I will buy u a drink, Hone

The selected Model is based on Random Forest With Kernel and the Prediction is : ['negative']

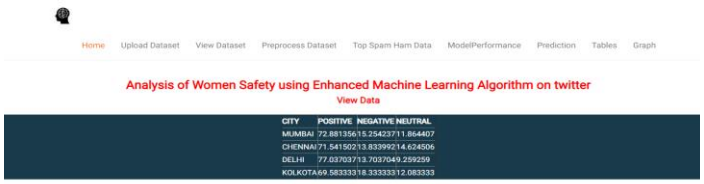


Analysis of Women Safety using Enhanced Machine Learning Algorithm

TWEET: Ok lar... Joking wif u oni

The selected Model is based on Random Forest With Kernel and the Prediction is : ['positive']

The "Tables" page displays city-wise sentiment analysis results, showing the percentage of positive, negative, and neutral tweets.



Analysis of Women Safety using Enhanced Machine Learning Algorithm on twitter

View Data

| CITY | POSITIVE | NEGATIVE | NEUTRAL |
|---|---|---|---|
| MUMBAI | 72.881356 | 15.254237 | 11.864407 |
| CHENNAI | 71.541502 | 13.833992 | 14.624506 |
| DELHI | 77.037037 | 13.703704 | 9.259259 |
| KOLKOTA | 69.583333 | 18.333333 | 12.083333 |

Finally The "Graph" page visually compares the performance metrics (accuracy, precision, and recall) of various machine learning models using a bar chart.

## VIII. CONCLUSION

Women and girls across the country face various forms of violence and harassment in public spaces, from stalking to sexual assault. This application aims to foster a sense of collective responsibility within Indian society to enhance women's safety in their environment. By analyzing tweets about women's safety in India often containing text, images, and comments we can predict sentiment and classify the nature of these tweets. Interpreting these insights enables us to take meaningful steps toward improving women's safety.

## REFERENCES

[1] Agarwal, A., Biadsy, F., & McKeown, K. R. (2009). Contextual Phrase-Level Polarity Analysis Using Lexical Affect Scoring and Syntactic N-grams. 12th Conference of the European Chapter of the Association for Computational Linguistics.

[2] Barbosa, L., & Feng, J. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy Data. 23rd International Conference on Computational Linguistics.

[3] Bermingham, A., & Smeaton, A. F. (2010). Classifying Sentiment in Microblogs: Is Brevity an Advantage? 19th ACM International Conference on Information and Knowledge Management.

[4] Gamon, M. (2004). Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. 20th International Conference on Computational Linguistics.

[5] Kim, S.-M., & Hovy, E. (2004). Determining the Sentiment of Opinions. 20th International Conference on Computational Linguistics.

[6] Klein, D., & Manning, C. D. (2003). Accurate Unlexicalized Parsing. 41st Annual Meeting of the Association for Computational Linguistics.

[7] Charniak, E., & Johnson, M. (2005). Coarse-to-Fine N-Best Parsing and MaxEnt Discriminative Reranking. 43rd Annual Meeting of the Association for Computational Linguistics.

[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.

[9] Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment Analysis on Twitter Data. International Journal of Innovative Research in Advanced Engineering (IJIRAE).

[10] Mamgain, N., Mehta, E., Mittal, A., & Bhatt, G. (2016). Sentiment Analysis of Top Colleges in India Using Twitter Data. International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT).