ISSN 2277-2685
IJESR/Jul-Sep. 2024/ Vol-14/Issue-3/1-14
Aarapelli Sree Bhavani *et. al.,* / International Journal of Engineering & Science Research

UNVEILING THE WOMEN'S SAFETY IN INDIAN CITIES THROUGH MACHINE LEARNING ON TWEETS

1 Aarapelli Sree Bhavani, University College of Engineering, Science and Technology, JNTUH Hyderabad

2 Dr. M Dhanalakshmi, Professor of IT, University College of Engineering, Science and Technology, JNTUH Hyderabad

Abstract: From rape to following, ladies and young ladies in Indian urban communities frequently experience different kinds of viciousness and badgering out in the open spots. The current review examines the capability of web-based entertainment stages, explicitly Instagram, Facebook, and Twitter, to work on the security of ladies. It investigates the manners by which these stages could assist Indians with fostering a feeling of social obligation to place ladies' wellbeing first in their networks. This study expects to tackle the force of virtual entertainment to decidedly influence Indian youth culture by concentrating on tweets and other web-based entertainment data connecting with ladies' wellbeing in Indian urban communities. It upholds the spread of messages that instruct and bring issues to light about ladies' wellbeing and calls for intense disciplines for harassers. Twitter is an imperative discussion for ladies to share their encounters and stresses over security in open regions due to its enormous following and hashtag developments. This study features the meaning of using ML approaches for the effective examination of web-based entertainment information, subsequently supporting endeavors to ensure more secure settings for ladies in Indian urban communities.

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

I. INTRODUCTION

A few types of savagery and provocation are especially forceful, for example, gazing and offering hostile comments, and these ways of behaving are normally acknowledged as common parts of metropolitan life. Various examination studies have been done in Indian urban areas, where ladies have announced encountering practically identical types of lewd behavior and overly critical comments from unidentified people. As per an overview that was completed in India's most crowded cities, including Delhi, Mumbai, and Pune, 60% of ladies report feeling risky while involving public transportation or leaving their homes for work. Since ladies reserve the privilege to the city, they are allowed to go at whatever point and any place they like, even to instructive establishments. Notwithstanding, in light of the fact that such countless unidentified Eyes are body-disgracing and bothering these ladies, they feel uncertain in regions like shopping centers and retail outlets while going to their work environments. The essential driver of young lady badgering is either wellbeing concerns or an absence of clear repercussions in ladies' lives. A few young ladies have encountered badgering from their neighbors while they were strolling to school, or there might have been an absence of security that made little kids unfortunate. These young ladies endure all through their lives because of that one occurrence wherein they were caused to do something despite their desire to the contrary to or were exposed to lewd behavior by a neighbor or an obscure person. Urban communities that focus on ladies' privileges to impact the city unafraid of savagery or inappropriate behavior are known for having the most secure approaches for ladies. It is the obligation of society to exactly characterize the requirement for security of ladies and to recognize that ladies and young ladies have similar right to somewhere safe in the city as men do, rather than putting limits on them as society ordinarily does. The names of people and ladies who stand in opposition to lewd behavior and exploitative way of behaving of guys in Indian urban communities that makes it hard for them to move uninhibitedly are likewise remembered for the examination of the assortment of texts from Twitter. The informational index about ladies' wellbeing in Indian culture that was gotten through Twitter was handled through ML calculations to smooth the information by dispensing with zero qualities. Also, the information was examined utilizing Laplace and Doorman's hypothesis to foster an examination strategy and take out repetitive and retweeted information from the informational index, bringing about a unique and clear image of the wellbeing status of ladies in Indian culture.

II. LITERATURE SURVEY

Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams:

A classifier to foresee the context oriented extremity of emotional expressions in a sentence is introduced by Agarwal et al. [1]. Our technique utilizes lexical scoring that is stretched out through WordNet and created from the Dictionary of Affect in Language (DAL). This permits us to score by far most of words in our feedback consequently, dispensing with the requirement for manual marking. To represent the impact of setting, the creators join n-gram investigation with lexical score. The creators remove n-grams of parts from each expression by consolidating syntactic constituents with DAL evaluations. The extremity of each syntactic component in the expression is one more element utilized by the essayists. Our discoveries show a critical improvement contrasted with both a greater part class benchmark and a seriously difficult standard comprised of lexical n-grams.

Robust sentiment detection on twitter from biased and noisy data:

To consequently recognize sentiments in Twitter messages (tweets), Barbosa et al. [2] introduced a technique that views at the organization of words in tweets as well as specific parts of their composing style. Also, the creators utilize loud mark sources as our preparation set. A couple of feeling acknowledgment sites provided these uproarious marks in light of Twitter information. Our investigations show to the creators that our strategy is stronger and effective than earlier ones while managing slanted and loud information, the kind of information that these sources supply, in light of the fact that our highlights can catch a more dynamic portrayal of tweets.

Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis:

Gamon et al's. research [3] shows that mechanized opinion classification might be done even in the loud field of buyer criticism information. The creators exhibit how to prepare direct help vector machines with great grouping precision on information that posture issues for even a human annotator by consolidating highlight decrease with large component vectors. The creators likewise show that, fairly shockingly, order execution in this space is reliably improved by including profound phonetic examination qualities into an assortment of superficial word n-gram highlights.

Study of Twitter sentiment analysis using machine learning algorithms on Python:

The concentrate by Gamon et al. [3] shows that robotized assessment arrangement is conceivable even in the uproarious area of client grumbling information. By consolidating feature decrease with large part vectors, the creators tell the best way to make direct assistance vector machines with magnificent gathering accuracy on information that presents issues for even a human annotator. The creators likewise show that, fairly shockingly, adding profound phonetic examination elements to an assortment of superficial word n-gram features reliably further develops request execution in this space.

III. EXISTING MODEL

This study researches how virtual entertainment locales like Facebook, Instagram, and Twitter could uphold ladies' wellbeing in Indian urban communities. It features how significant these gatherings are in spreading mindfulness, empowering responsibility in Indian culture, and pushing for more grounded regulations against provocation. The review centers around the manners by which ladies might advocate for more secure public regions and offer their encounters utilizing hashtags and messages on Twitter.

Disadvantages:

1. Has restricted convenience in disconnected conditions since it relies upon web accessibility to get to and examine tweets.

2. Opinion examination may not utilize refined normal language handling methods, which could bring about less shrewd outcomes.

IV. METHODOLOGY

A. Proposed Model

With the utilization of a critical number of Twitter information —, for example, time tweets, dates, and sources — the as of late recommended Ladies Wellbeing Framework means to fundamentally improve how much text information by including extra examples. High level text handling will be applied to kill pointless words and induce missing qualities with the goal that the attention might be on appropriate data. Choosing the ideal model to conjecture and assess security related tweets will be made simpler by testing a few ML calculations on the latest information. A simple to-utilize online application will be created with a natural UI, ongoing information refreshes, and clever illustrations to empower clients to discover security related data quickly. Notwithstanding intensive information gathering and modern information investigation, the drive will expand ladies' wellbeing by thoroughly assessing models and offering a natural application.

Advantages:

1. Equipped for taking care of tweets disconnected, ensuring information openness even without a trace of a web association.

2. Gives further experiences into public feeling using complex NLP advances like NLTK opinion investigation.

B. Modules

• Upload dataset: We'll transfer the Twitter dataset utilizing this module.

• Dataset cleaning: This module will recognize any unfilled qualities in the dataset and supplant them with the mean or zero qualities.

• Train and Test Split: We will separate the dataset into equal parts, alluded to as preparing and testing, utilizing this module. A classifier is prepared on 70% of the dataset utilizing ML techniques, and its forecast accuracy is tried on 30% of the dataset.

• Run Models: This module will be utilized to prepare different models, for example, SVC, Random Forest, Decision Tree, and Logistic Regression classifier, utilizing 70% of the information partitioned and 30% of the information used to decide the model's exhibition.

• Predict Result: Train each classifier that was picked, estimate results, and evaluate adequacy.

C. System Architecture:

The's framework will probably frame the means associated with making a ladies security framework, beginning with information refinement, as recognized in the Architectural Meanderings. The cycles incorporate, among others, highlight determination to separate relevant data, text preprocessing to tidy up the text information, and making of a TF-IDF lattice to show the element's importance in the report. The perforated queen (in the information) was alloted to preparing and testing after information planning. This is a subconscious delineation of the contention among stress and mystery while educating a youngster — for example, by using the children's actual names. The picked

ISSN 2277-2685
IJESR/Jul-Sep. 2024/ Vol-14/Issue-3/1-14
Aarapelli Sree Bhavani *et. al.,* / International Journal of Engineering & Science Research

model is presently being planned and fabricated. Once finished, it is carried out, and a report is displayed on the landing page. It is the simplest method and allows for greater user comfort.

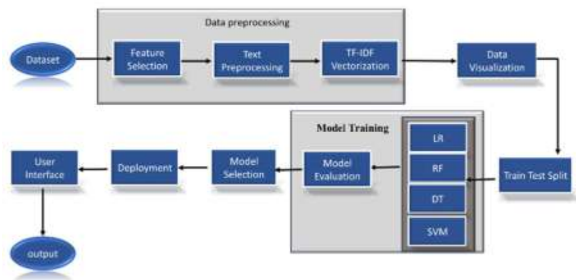

Fig 1: System Architecture

D. Dataset:

All 28,000+ tweets sent on Twitter/X all through the development are remembered for the "MeToo Tweets" dataset, which was taken from Kaggle. The tweetid, made (timestamp), text (tweet content), retweets, top picks, source.r (wellspring of the tweet), hashtags, num.emojis (number of emoticons), and emoji_names are a few additional properties that are given in this effectively open dataset from Kaggle. With 28,360 rows and 9 columns altogether, the dataset offers a lot of room for the examination of public discussion around the Me Too development. This enormous dataset has made it feasible for scholastics to examine the examples, patterns, and feelings that arose all through the Me Too era. Subsequently, new instructive understandings about the social impact and extent of the Me Too development on different virtual entertainment stages have arisen.

E. Data Processing:

The text property was chosen as the sole component for assessment during the pre-handling stage. Refined procedures for text cleaning were applied, remembering filling for clear qualities, disposing of alphanumeric words, changing the text's case, adding spaces to newlines, and wiping out non-ASCII characters. From that point onward, the cleaned text was encoded into mathematical qualities utilizing TF-IDF vectorization, which permitted the model to extricate includes and investigate them for preparing. These strategies ensure that the information is uncontaminated and ready for extra review or model preparation.

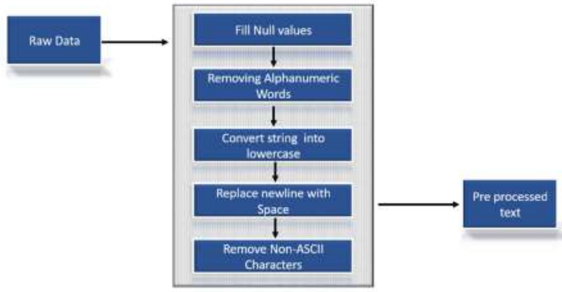

Fig 2: Text Preprocessing Architecture

F. Algorithms

Support Vector Classifier:

SUPPORT VECTOR CLASSIFIER This characterization calculation upholds both direct and non-straight relapses, as its name would infer. The Support Vector Machine is the establishment whereupon this method works. SVM is a classifier that is utilized to foresee discrete downright marks,whereas SVC is not.

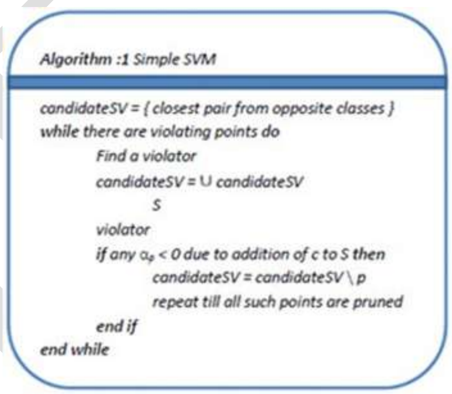

Fig 3: SVM Pseudo Code

While we limit the blunder rate in essential relapse, in Support Vector Classifier (SVC), we fit the mistakes into a stretch known as the ε-tube. Inside that edge, SVC decides the ideal worth. Significant thoughts incorporate the accompanying: the limit lines, which structure an edge around the hyperplane; the help vectors, which are the outrageous information focuses that characterize the hyperplane; the hyperplane, what partitions the classes in a higher aspect and predicts the objective qualities; and the piece, which maps the information focuses to a higher aspect (e.g., Sigmoidal, Polynomial, Gaussian). In SVM order, the help vectors unbendingly characterize the hyperplane, however in SVC, as numerous data of interest as practical are fitted without conflicting with the edge.

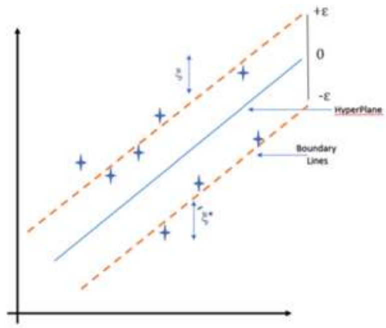

Fig 4: SVC Hyper plane Diagram

Decision Tree:

One famous ML approach that handles both relapse and order issues is the decision tree. It works by isolating the dataset into more modest gatherings and making a decision tree-like model by using the main trademark as a node. To make a tree, we initially pick the main component utilizing entropy, difference decrease, or Gini contamination measures. Then, we partition the information utilizing recursive separation. The class mark or nonstop objective worth is the terminal node, the branch is the experimental outcome, and the inward nodes are the divided focuses on a trait. This cycle go on until the end necessities are fulfilled, for example, the greatest profundity or the necessary least number of tests per leaf. It is a flexible device for prescient examination since a fathomable model can be built freely like a creation task for both mathematical and straight out information.

```
calculate overall entropy
for each value of each attribute
      calculate the entropy value and attribute gain
      select the winner attribute
         for each value on winner attribute
            create a new branch
            if entropy value equals zero then
               create a leaf
            else
               create a new child node
            end if
         end for
      end select
      add father node
end for
```

Fig 5: Decision Tree Pseudo Code

Random Forest:

Among the administered learning strategies is the notable machine learning algorithm Random Forest. It very well might be applied to ML issues including both order and relapse. Its establishment is the possibility of troupe realizing, which is the demonstration of combining a few classifiers to settle a difficult issue and upgrade the model's usefulness. As indicated by its name, "Random Forest is a classifier that contains various choice trees on different subsets of the given dataset and takes the normal to work on the prescient precision of that dataset." Rather of relying upon a solitary decision tree, the random forest gauges the result in view of the greater part vote of projections from each tree. Since there are more trees in the forest, accuracy is higher and overfitting is avoided.

```
Input: N - Quantitative amount of bootstrap samples
           M - Total number of features
           m - Sample size
           k - Next node
Output: A Random Forest (RF)
Steps:
1. Creates N bootstrap samples from the dataset.
2. Every node (sample) takings a feature randomly of size m where m<M.
3. Builds a split for the m features selected in Step 2 and detects the k node by using
   the best split point.
4. Split the tree iteratively until one leaf node is attained and the tree remains
completed.
5. The algorithm is trained on each bootstrapped independently.
6. Using trees classification voting predicted data is collected from the trained trees
(n).
7. The final RF model is build using the peak voted features.
8. return RF
End.
```

Fig 6: Random Forest Pseudo Code

The below diagram explains the working of the Random Forest algorithm:

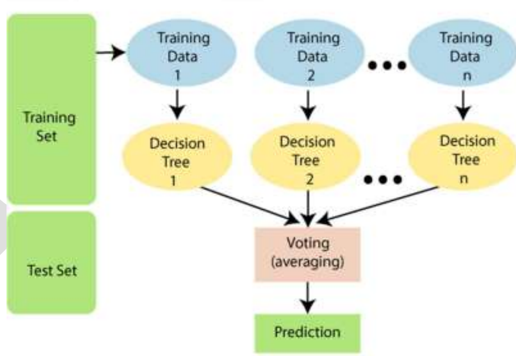

Fig 7: Random Forest Diagram

Logistic Regression:

While managing multi-class issues (i.e., multiple classes), calculated relapse utilizes different methods, like OvR (One-versus Rest). OvR fits double calculated relapse classifiers for every one of the three classes autonomously at three classes. Each sophisticate picks whether to have a place with one class or the consolidated two. The class with the most noteworthy probability across the three classifiers is chosen as the last reaction in the forecast cycle. On the other hand, Softmax relapse utilizes a softmax capability to make sense of the model, it are 1 to imply that all anticipated probabilities. This permits Softmax relapse to stretch out strategic relapse to numerous classes. This is significant when the strategy returns the general likelihood dispersion across all classes for each info case and the classes are exhaustive and idiosyncratic.

Input:

- *Training algorithm L (logistic regression)*
- *Sample matrix X*
- *Labels vector* $y = [1, \dots K]$
- *Initial regresor parameters vector* θ_i

Main:

For i=1:K

Create a new binary vector y_i *for each label where* $y_i = 1$ *if it belong to the label and* $y_i = 0$ *if it does not belong.*

Apply L to X to find θ_i

Output:

θ_i *Parameters vector for each regresor*

Fig 8: Logistic Regression Pseudo Code

V. EXPERIMENTAL RESULTS

a) Model Comparison:

The reference chart shows four particular models alongside instances of each with regards to testing and preparing exactness. The right accuracy rate is displayed on the y-pivot, which goes from 0% to 100 percent. There are five sets of bars on the outline's x-pivot, every one of which addresses an unmistakable sort of model. The light blue groups address test accuracy, while the red bars address the preparation focuses. Review for the Strategic Relapse model is 92.42% during preparing and 88.05% during testing. With 75.28% of the preparation stage and 93.15% of the testing stage finished, the SVM model has likewise demonstrated fruitful. The Decision Tree model has the best preparation accuracy of 99.94%, however its trying exactness of 80.39% is very poor. On the other hand, the Random Forest model has a generally low relationship with preparing (44.58%), which is likewise reflected in a comparatively low testing score (45.97%). This proposes that out of the relative multitude of models, the Calculated Relapse model is the most conventional.

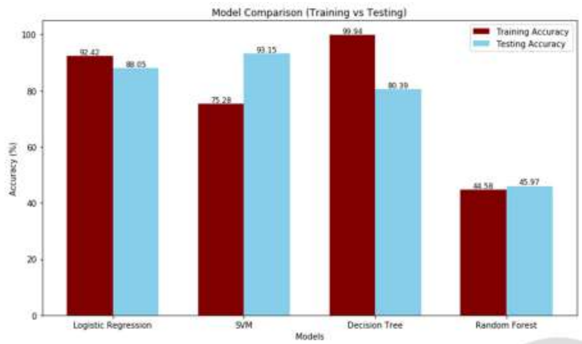

Fig 9: Comparison of different Models

b) Logistic Regression Confusion Matrix:

The Logistic Regression model's disarray grid gives subtleties on the model's order conduct as for three states: negative, neutral, and positive. It demonstrates the way that the model can gauge 9,735 positive cases, 7,864 impartial cases, and 3,567 negative cases. There's likewise an opportunity that 412 negatives turned unbiased, 543 negatives turned positive, 69 neutrals turned negative, 189 neutrals turned positive, 200 up-sides turned negative, and 324 negatives turned positive were all wrong. Obviously the initial two rating classes are the ones that were generally inaccurately expected, and the positive class isn't quite so vigorously designated as the classes that preceded it. While the model truly does well generally, it battles a piece to differentiate between the positive and negative classes.

Fig 10: Confusion Matrix

c) Evaluation Metrics:

The accompanying exhibition numbers are accommodated the Logistic Regression algorithm: Accuracy 88.0%, Precision 88.1%, Recall 88.0%, and F1 Score 87.7%. These estimations exhibit that the calculated relapse model might yield a serious level of exactness and trustworthiness, as well as an extraordinary F1 score and general rightness, making it the ideal option eventually.

Evaluation Metric	Values
Accuracy	88.0%
Precision	88.1%
Recall	88.0%
F1 Score	87.7%

Fig 11: Evaluation Metrics Table

d) Application Results

Fig 12: Home Page

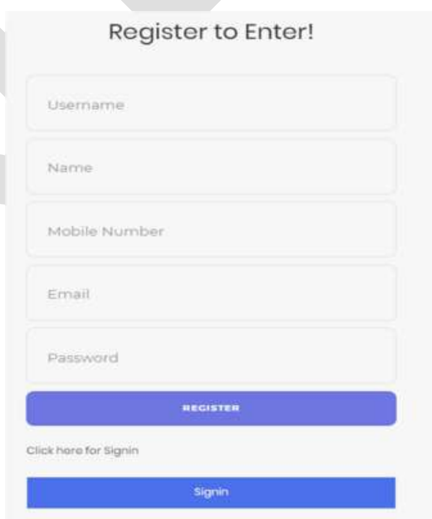

Fig 13: Registration Page

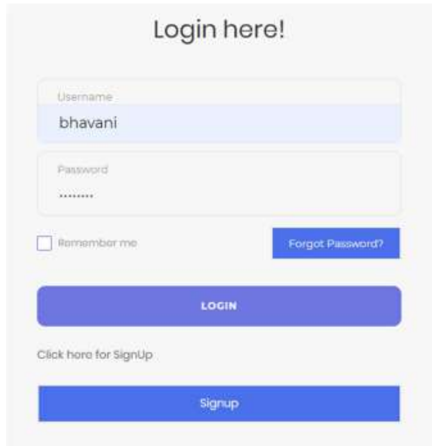

Fig 14: Login Page

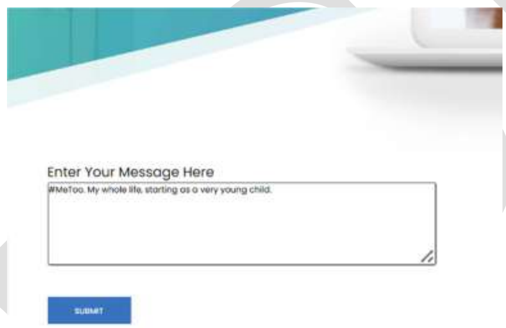

Fig 15: Input Screen

Fig 16: Output Screen

VI. CONCLUSION

This study shows how different ML calculations and natural language processing (NLP) might be successfully used to screen and assess the monstrous measure of Twitter information sources that are created consistently. With a testing accuracy of 88%, Logistic Regression arose as the top entertainer among the models subsequent to representing different elements. This surprising accuracy features its relevance to feeling investigation issues. Using the web-based application, clients might enter message information into a textbox to get the feeling examination discoveries, which can be either good, unbiased, or pessimistic. The vigor and effectiveness of calculated relapse continuously opinion examination on huge web-based entertainment informational collections are exhibited by serious areas of strength for its. As a rule, this is the best delineation of how ML and natural language processing (NLP) can change natural online entertainment into quick information that might illuminate wise decisions.

VII. FUTURE ENHANCEMENT

How ladies' security frameworks are further developed in the future through the utilization of innovation like strategic relapse will make a huge difference. Through the assortment of information from Twitter and other web-based entertainment stages, the framework can get an assortment of continuous client inputs, for example, tweets with respect to occasions or wellbeing concerns. By picturing this information, protection security activities will be made conceivable as well as bits of knowledge into examples and areas of interest. Upgrades continuously information investigation will take into account quick response instruments by means of versatile applications and essentially further develop the forecast models. In light of their area or virtual entertainment conduct, this will tell clients of any potential dangers. Moreover, adding opinion investigation and regular language handling to the framework will work on its capacity to perceive and deal with trouble signals. In light of everything, these advancements will work on the responsiveness, proactiveness, and continuous customization of ladies' security applications to the always changing social scene and client wellbeing concerns.

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.

[11] Deepak Kumar, Alpana Kakkar, Anand Kumar Dubey, "Pattern Analysis in Academics for Mentor Systems", 2021 International Conference on Computational Performance Evaluation (ComPE), pp.151-154, 2021.

[12] Priyanka Kohli, Kawaljeet Singh, "Analysis of Woman Safety Parameters in Smart and Non-Smart Cities", 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp.1-5, 2021.

[13] Diksha Rana, Deepak Kumar, Maya Kumari, Rina Kumari, "Assessing the Impact of Delhi Metro Network Towards Urbanisation of Delhi-NCR", Geospatial Technology for Landscape and Environmental Management, pp.333, 2022.

[14] Emma Xue, Bolun Zhang, Qiaoru Zhang, Ran Hu, Jielin Jiang, Nian Liu, Yingdong Peng, Ziqian Li, Judith Logan, , 2023.