

## **Analysis of Women Safety in Indian cities using Machine Learning on Tweets**

*M.TEJASHWINI, VENNA LAKSHMI PRASANNA, SUBBARA PRAVALLIKA, SHAIK AFROZ,  
RACHAGOLLA RAGHUVARAN YADAV, MADIGA DHANUNJAYA*

*DEPT OF CSE*

*KRISHNACHAITANYA INSTITUTE OF TECHNOLOGY & SCIENCES*

### **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?

### **I.INTRODUCTION**

Twitter in this modern era has emerged as a ultimate microblogging social network consisting over hundred million users and generate over five hundred million messages known as 'Tweets' every day. Twitter with such a massive audience has magnetized users to emit their perspective and judgemental about every existing issue and topic of internet, therefore twitter is an informative source for all the zones like institutions, companies and organizations. On the twitter, users will share their opinions and perspective in the tweets section. This tweet can only

contain 140 characters, thus making the users to compact their messages with the help of abbreviations, slang, shot forms, emoticons, etc. In addition to this, many people express their opinions by using polysemy and sarcasm also. Hence twitter language can be termed as the unstructured. From the tweet, the sentiment behind the message is extracted. This extraction is done by using the sentimental analysis procedure. Results of the sentimental analysis can be used in many areas like sentiments regarding a particular brand or release of a product, analyzing public opinions on the government policies, people thoughts on women, etc. In order to perform classification of tweets and analyze the outcome, a lot of study has been done on the data obtained by the twitter. We also review some studies on machine learning in this paper and research on how to perform sentimental analysis using that domain on twitter data. The paper scope is restricted to machine learning algorithm and models.

Staring at women and passing comments can be certain types of violence and harassments and these practices, which are unacceptable, are usually normal especially on the part of urban life. Many researches that have been conducted in India shows that women have reported sexual harassment and other practices as stated above. Such studies have also shown that in popular metropolitan cities like Delhi, Pune, Chennai and Mumbai, most women feel they are unsafe when surrounded by unknown people. On social media, people can freely express what they feel about the Indian politics, society and many other thoughts. Similarly, women can also share their experiences if they have faced any violence or sexual harassment and this brings innocent people together in order to stand up against such incidents. From the analysis of tweets text collection obtained by the twitter, it includes names of people who has harassed the women and also names of women or innocent people who have stood against such violent acts or unethical behaviour of

men and thus making them uncomfortable to walk freely in public.

The data set of the tweet will be used to process the machine learning algorithms and models. This algorithm will perform smoothening the tweet data by eliminating zero values. Using Laplace and porter's theory, a method is developed in order to analyze the tweet data and remove redundant information from the data set. Huge numbers of people have been attracted to social media platform such as Twitter, Facebook, Instagram. People express their sentiments about society, politics, women, etc via the text messages, emoticons and hash-tags through such platforms. There are some methods of sentiment that can be classified like machine leaning based and lexicon based learning

## II. EXISTING SYSTEM:

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.

## III. PROPOSED SYSTEM:

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 retweet 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

## IV. SYSTEM ARCHITECTURE:

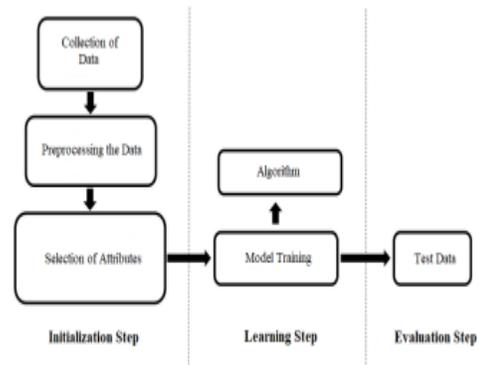


Fig -1: Process of Analysis

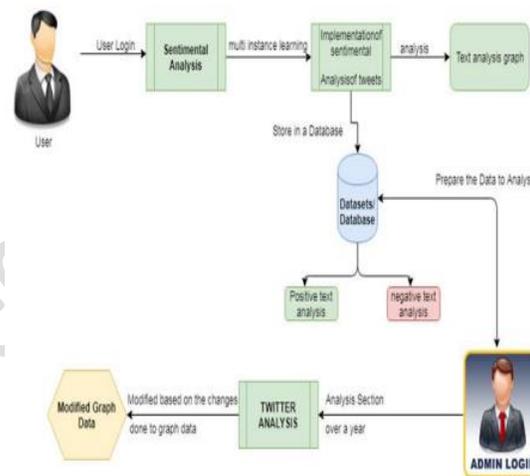


Fig -2: Architecture

## V. MODULES:

**1) Data extraction:** First step involved in analysis of sentiment is the collection of information from the social network website like twitter. This helps in extracting the tweet message but this message also includes extra data like tweets likes, dislikes and comments.

**2) Text Cleaning:** Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.

**3) Sentiment Analysis:** After the classifier cleans the dataset, the data is ready for the sentimental analysis process. Machine learning and Lexicon based learning and Hybrid learning are some of the approaches of sentimental analysis. There are also some other approaches such as Nero Linguistic Programming and Natural Language Processing. Training the dataset and then testing that trained dataset involves in machine learning approach. Training data and Testing data are useful for the classifier to perform the algorithm. Maximum Entropy, Naives Bayes classification, Bayesian Networks and Network Support Vector Machine are some of the algorithm which can be used to train the classifier. Testing data is used to identify the efficiency of the sentiment classifier. In case of Lexicon based leaning, training dataset is not used. This approach uses a built-in dictionary in which words associated with sentiments of human are present. The third approach, which is the Hybrid learning, combines both machine leaning approach and lexicon learning approach in order to improve the performance of classifier.

**4) Sentiment Classification:** At this step, the dataset is ready for the classification. Each and every sentence of the tweet will be examined and opinion will be formed accordingly for subjectivity. Subjective expression sentences are retained and those of objective expression sentences are rejected. Techniques like Unigrams, Negation, Lemmas and so on are used at different levels of sentimental analysis. Sentiments can be distinguished broadly into two groups – Positive and Negative. At this point of sentimental analysis, each of the subjective sentences which will be retained are classified into good, bad or like, dislike or positive and negative.

**5) Output Presentation:** To generate useful and meaningful information out of the raw data, sentimental analysis plays vital role. Once the algorithm is completed, the outcome of the analysis can be visualized by creating different types of graphs. Bar graphs, Time series and Pie charts are some of the examples which can be used to display the output. To measure the sentiment of the tweets in terms of Positive and Negative, Bar graphs can be used. Similarly, to measure in terms of likes, dislikes, average length of tweet for a certain period, Time

series can be used. To obtain the initial source of the tweet, pie charts can be used.

## VI. SYSTEM SPECIFICATION:

### HARDWARE REQUIREMENTS:

- ❖ **System** : Pentium IV 2.4 GHz.
- ❖ **Hard Disk** : 40 GB.
- ❖ **Floppy Drive** : 1.44 Mb.
- ❖ **Monitor** : 14' Colour Monitor.
- ❖ **Mouse** : Optical Mouse.
- ❖ **Ram** : 512 Mb.

### SOFTWARE REQUIREMENTS:

- ❖ **Operating system** : Windows 7 Ultimate.
- ❖ **Coding Language** : Python. anaconda
- ❖ **Front-End** : html

## VII. 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.

## VIII. REFERENCES

- [1] Agarwal, Apoorv, Fadi Biadisy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12<sup>th</sup> 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] Mangain, 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.