International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Analysis of Women Safety using Machine Learning on Tweets

Ruchira Kadam [1], J. C. Pasalkar [2]

[1]Student, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune, Maharashtra, India

[2]Associate Professor, Department of Information Technology, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune, Maharashtra, India

ABSTRACT

In this period, as we realize there are various preceding instances in the world regarding women harassment that's beginning from stalking leading to abusive harassments which include acid assaults, rape cases, obscenity, pornography and so on. With such issues, an evaluation of women safety in Indian cities has been proposed the usage of system mastering on Tweets. The Sentimental analysis is taken as the primary concept and is done through gadget getting to know, taking an enter from tweets to ensure protection. This sentimental evaluation on tweets helps in growing focus amongst human beings. As we recognize Twitter and Instagram are accountable to unfold statistics far and extensive among the people across the globe, this helps the women in expressing her emotions to the world. we've taken Twitter as an crucial useful resource because it presents text, audio, video message and are smooth to handle. this may assist our studies to overcome the emotions of humans around us.

INTRODUCTION

There are certain kinds of harassment and Violence that are very aggressive along with staring and passing remarks and those unacceptable practices are commonly seen as a normal part of the urban life. There are instances while the harassment of girls turned into executed by means of their neighbours while they have been on the way to highschool or there has been a lack of safety that created a sense of worry inside the minds of small women who at some point of their lifetime go through because of that one example that passed off in their lives where they had been pressured to do something unacceptable or turned into sexually harassed by using one of their own neighbor or some other unknown character. safest cities method girls protection from a perspective of ladies rights to the affect the metropolis with out fear of violence or sexual harassment. instead of enforcing regulations on women that society normally imposes it's far the responsibility of society to imprecise the need of protection of women and additionally recognizes that women and ladies actually have a proper identical as guys must be secure inside the metropolis.

analysis of twitter texts collection additionally includes the call of humans and name of women who stand up in opposition to sexual harassment and unethical behaviour of men in Indian cities which cause them to uncomfortable to walk freely. The information set that turned into acquired via Twitter about the repute of ladies safety in Indian society become for the processed via machine mastering algorithms for the purpose of smoothening the data through casting off zero values and the usage of Laplace and porter's concept is to developer technique of analyzation of statistics and remove retweet and redundant information from the facts set that is acquired so that a clean and unique view of safety repute of girls in Indian society is acquired.

Twitter in this contemporary generation has emerged as a ultimate microblogging social community consisting over hundred million customers and generate over five hundred million messages known as 'Tweets' every day. Twitter with this kind of big target audience has magnetized customers to emit their perspective and judgemental about each present issue and topic of net, consequently twitter is an informative supply for all the zones like institutions, organizations and corporations. on the twitter, users will proportion their reviews and attitude in the tweets phase. This tweet can most effective comprise one hundred forty characters, as a result making the users to compact their messages with the assist of abbreviations, slang, shot forms, emoticons, and so on. in addition to this, many humans express their opinions with the aid of the use of polysemy and sarcasm additionally.

RELATED WORK

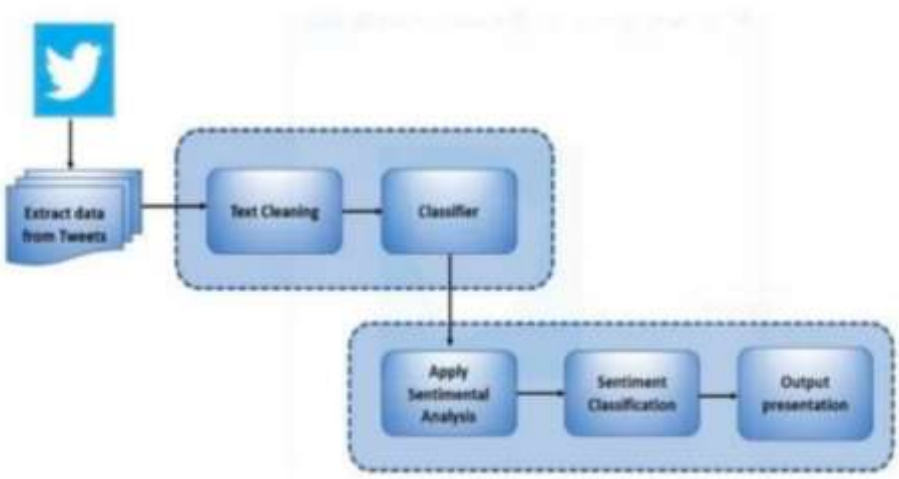
Sentiment Analysis (SA) is a subject of study that investigates people's sentiments, views, assessments, appraisals, attitudes and emotions in the direction of entities such as individuals, services, organizations, issues, products, topics and their characteristics. It is also known as opinion mining, sentiment mining, subjectivity analysis, review mining, opinion extraction, emotion analysis, etc.

Sentiment methods	Classification	Pros and Cons
Lexicon based	Dictionary based. Corpus based. Ensemble approaches.	Pros: Best for domain reliant on, larger-term coverage. Cons: Only Finite number of words in the lexicons

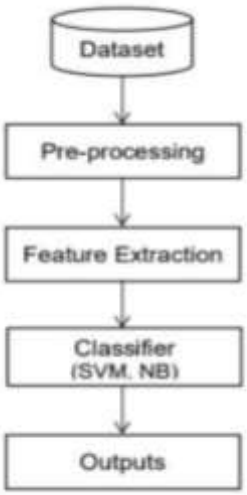
Machine learning Based	Support vector machines. Bayesian networks. Naïve Bayes. Random forest.	Pros: Capacity to adjust and make prepared models for explicit purposes and settings. Cons: Low relevance for new information, since it is important of marked information.
Hybrid based	Lexicon and machine learning based.	Pros: High exactness of new information. Slant vocabulary developed utilizing public assets for assumption discovery. Notion words as highlights in the AI technique. Cons:Noisy data

SENTIMENTAL ANALYSIS

Sentimental analysis is the manner of drawing out the viewpoint behind the sentences or statements. it's miles manner used to reap the viewpoints from twitter. those viewpoints help in reaching sentiment category. Viewpoints vary from person to person, so it is critical to apprehend what all people is attempting to deliver. The man or woman performing the sentimental evaluation must decide the classifications to be made primarily based on the information. because it's far an essential element to determine the efficiency of the set of rules. based totally on the twitter statistics we can classify it as wonderful, poor, or neutral. There are two kinds of tactics: system learning and Lexicon getting to know. machine getting to know consists of the procedure of extraction of functions, programming model education the use of dataset of functions. On the opposite and Lexicon gaining knowledge of based method uses the vocabulary and scoring approach to locate viewpoints. on this paper we're the usage of gadget gaining knowledge of method. statistics extraction, text cleansing, sentimental analysis, type, output presentation are the principal steps concerned



METHODOLOGY



In the Sentiment Analysis the following steps are major to identify the positive, negative or neutral of the twitter post.

They are:

i. Collecting the Dataset.

ii. Pre-Processing the Dataset.

iii. Feature Extraction.

iv. Apply Classifier.

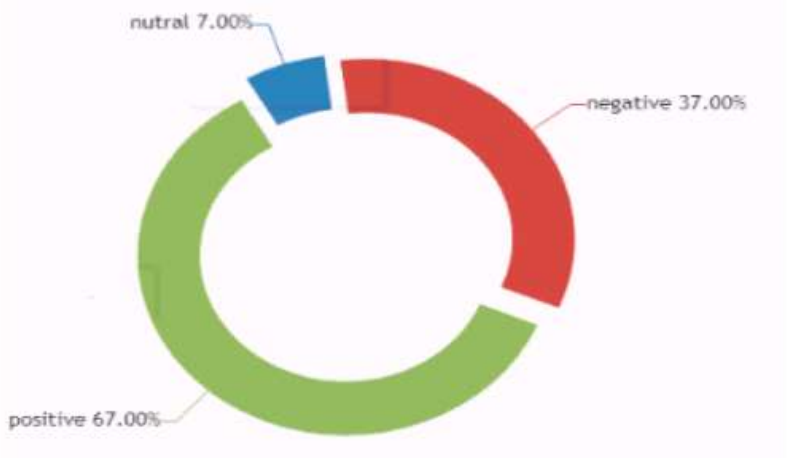
- **i.** **Data Collection:** The data is gathered from the twitter using API. Application program interface (API) is utilized to gather the information. Twitter website is a source which consists of users tweets.
- **ii.** **Data Preparation:** In Data Pre-processing removing noisy, unrelated data, inconsistent and incompletedata from the dataset. Generally, in twitter we have to remove URLs, special characters, retweets, hash tags.
- **iii.** **Feature Extraction:** In this work, we used Bag of Words to extract features from text documents. After extraction, these features used for training machine learning algorithms. It makes a jargon of the apparent multitude of novel words happening in all the reports in the preparation set. Bag of words features containing term frequencies of each word in each document, i.e. the number of occurrence and not sequence or order of words matters. This can be done by CountVectorizer method in Python
- **iv.** **Classification:** A classification problem is applied if the output variable is a label or category, such as “Rainy” or “Sunny” or “disease” and “no disease” or in our work “Positive” or “Negative”

In the twitter information base, it stores the person information which include new tweets, re-tweets and tweet rating. The tweets which might be abusive to women are monitored and confirmed based totally on the user statistics. consumer tweets are used to provide protection to girls with the assist of sentimental analysis and will be stored within the facts base. To carry out the evaluation the admin makes use of the gathered dataset. inside the sentimental analysis the initial input called as application tweet is constructed by means of each user filter and can be saved inside the information base. Admin clear out is used to check whether the tweets is abusive to women or not. There are two varieties of keywords: high quality and poor key phrases. The advantageous key phrases are the ones which disrespect the girls, and the poor keywords are the ones which are not abusive to girls.

There may be many code phrases or keywords stored inside the twitter database. while the user applies sentimental analysis each and each word within the tweet of the consumer can be as compared with the code words stored in the statistics base. If the words within the tweet fit nice code words, it will be stored under advantageous code word within the statistics base (which can be abusive to women). in the equal way if it suits the terrible code words, it will likely be stored under terrible code words in the database (which aren't abusive to women). This class facilitates us inside the analysis. So, there will be two forms of sentimental analysis superb and terrible. fine analysis gives us the tweets which might be harmful to girls. similarly negative sentimental evaluation gives us the tweets that are smooth and not abusive to women. At this degree we can get the tweets in addition to person information.

RESULT

In this work, we took Twitter API database as input database, after that Pre-process the data set(remove incomplete and noisy data) then apply feature extraction method as bagging of words and finally used Naïve Bayes classification. In this work, we used python language to develop the system. The following figure is the result of the given Twitter API dataset Tweets Accuracy Percentage.



CONCLUSION AND FUTURE SCOPE

In the course of the research paper we've mentioned approximately various device getting to know algorithms that may assist us to organize and analyze the huge quantity of Twitter facts obtained inclusive of thousands and thousands of tweets and text messages shared each day. these machine getting to

know algorithms are very effective and beneficial. via the evaluation we've achieved the usage of gadget getting to know we can without difficulty discover the intensions of the humans ahead. device gaining knowledge of algorithm can manage the huge set of data successfully as well as effectively. this may absolutely help ladies to shield herself in lots of situations. This algorithm works properly in lots of structures. considering, twitter is the most standard platform it turns into smooth to spread facts and create recognition amongst humans regarding girls protection. If the intensions are terrible, then we will punish such human beings in order that we will avoid the probabilities of the assault on the girls.

For the future development, we can stretch out to apply those device gaining knowledge of strategies on numerous internet-based media ranges like face e book and imperative likewise since in our work simply twitter is idea of. present standpoint that is proposed can be incorporated with the twitter software interface to reach at bigger diploma and apply wistful exam on a large quantity of tweets to present extra health.

REFERENCES

[1] Eugene Charniak and additionally Mark Johnson. "Coarse-to-exceptional nbest parsing and MaxEnt discriminative reranking." Process of the 43rd annual assembly on association for computational linguistics. Organization for Computational Grammar, 2005.

[2] Gupta B, Negi M, Vishwakarma K, Rawat G &Badhani P (2017). "Research of Twitter sentiment evaluation making use of system mastering algorithms on Python." International Journal of Computer System Applications, one hundred sixty-five(nine) 0975-8887.

[3] Adam Bermingham and Alan F. Smeaton. "Classifying sentiment in microblogs: is brevity an advantage?." complaints of the 19th ACM global conference on information and expertise control. ACM, 2010.

[4] Mamgain N, Mehta E, Mittal A & Bhatt G (2016, March). "Belief evaluation of top schools in India making use of Twitter data." In Computational Strategies, in Details and Interaction Technologies (ICCTICT).

[5] Soo-Min Kim and Eduard Hovy. "determining the sentiment of evaluations." court cases of the twentieth international conference on Computational Linguistics. association for Computational Linguistics, 2004.

[6] Dan Klein as well as Christopher D. Manning. "Accurate unlexicalized parsing." Procedures of the forty first Yearly Meeting on Organization for Computational LinguisticsVolume 1. Association for Computational Linguistics, 2003

[7] Revathy K & Sathiyabhama B. (2013). A hybrid approach for supervised twitter sentiment classification. International Journal of Computer Science and Business Informatics.

[8] Sahayak V, Shete V & Pathan A (2015). “Sentiment analysis on twitter statistics.” international journal of revolutionary research in advanced Engineering (IJIRAE), 2(1), 178-183.