# A Mini Project on

# Analysis of women safety in Indian Cities using Machine learning on Tweets

A Project Report submitted in partial fulfilment of the degree of the

Bachelor of Technology in **COMPUTER SCIENCE AND ENGINEERING**

**(CSE)**

By

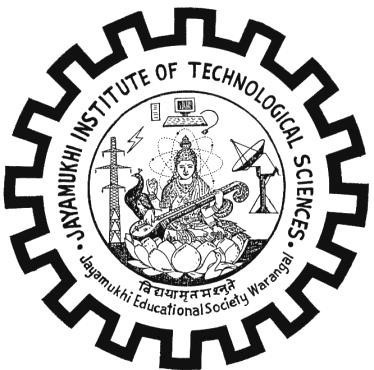
**SANIYA 21C41A0563**

**P. AKHILA 21C41A0542**

**P.SAI VISHNU 21C41A0518**

Under the Guidance of

**Mrs.G.JOSEMARY**



Department of **CSE**

**JAYAMUKHI INSTITUTE OF TECHNOLOGICAL SCIENCES**

(Autonomous, Affiliated to JNTUH & Accredited by NAAC with ‘A’ Grade) NARSAMPET,

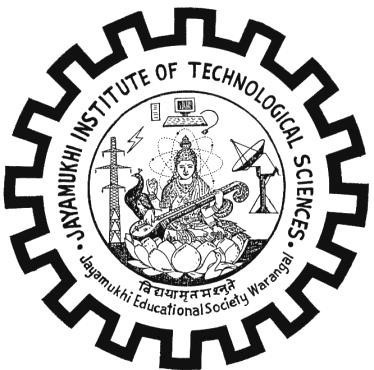
WARANGAL-506 332

2024

**JAYAMUKHI INSTITUTE OF TECHNOLOGICAL SCIENCES**

(Autonomous, Affiliated to JNTUH & Accredited by NAAC with ‘A’ Grade) NARSAMPET, WARANGAL – 506 332

**CERTIFICATE**



This is to certify that the project report entitled “**ANALYSIS OF WOMEN SAFETY IN INDIAN CITIES USING MACHINE LEARNING ON TWEETS**” is a bonafide work of the students **SANIYA, P.AKHILA,P.SAIVISHNU** roll No.s **21C41A0563,** **21C41A0542, 21C41A0518** submitted in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in CSE during the academic year 2024-2025.

### **Guide Head of the Department**

**Principal**

# ACKNOWLEDGEMENT

We express our gratitude and sincere thanks to **Prof. V. Janaki**, Principal**, Jayamukhi** **Institute of Technological Sciences** for her constant motivation and support.

We express our gratitude and sincere thanks **to Dr. G. Josemary**, Head of the Department CSE, , Jayamukhi Institute of Technological Sciences for her constant motivation and support.

We are grateful to our project guide **Mrs. G. Josemary**, professor for her valuable guidance and giving us the opportunity to do our work in a better manner.

We are also thankful to our management for providing all the facilities in completing the project.

**SANIYA 21C41A0563**

**P. AKHILA 21C41A0542**

**P.SAI VISHNU 21C41A0518**

# CONTENTS

**Page no**

1. [INTRODUCTION……………………………………………………….01](#_TOC_250024)
2. [LITERATURE SURVEY……………………………………………….02](#_TOC_250024)

3. [SENTIMENT ANALYSIS………………………………………………03](#_TOC_250019)

3.1[Existing system……………………………………………………03](#_TOC_250017)

* 1. Proposed system…………………………………………………... 03-04

4.SYSTEM REQUIRMENTS……………………………………………...05

4.1 [Software Requireme………………………………………………05](#_TOC_250017)

* 1. Hardware Requirements……………………………………………...05-06

**5.REQUIREMENTS………………………………………………………**06

5.1 Functional Requirements……………………………………………. 06

5.2 Non-Functional Requirements………………………………………..06-07

**6.SYSTEM STUDY…………………………………………………………**08

6.1 FEASIBILITY STUDY…………………………………………….…08

6.1.1 Economical Feasibility…………………………………………..08

6.1.2 Technical Feasibility……………………………………………08

6.1.3 Social Feasibility………………………………………………..08

**7.SYSTEM ARCHITECTURE………………………………………………**09-11

**8.UML DIAGRAMS………………………………………………………….**12

8.1 [Use Case Diagram…………………………………………………...12-13](#_TOC_250017)

8.2 Class Diagram…………………………………………………………...13-14

8.3 Object diagram……………………………………………………………………………………..14-15

8.4 State Diagram……………………………………………………………15-16

8.5 Activity Diagram…………………………………………………………………………………….17

8.6 Sequence Diagram……………………………………………………….18

8.7 Collaboration Diagram…………………………………………………..19

**9.SYSTEM IMPLEMENTATION……………………………………………**20-27

**10.CONCLUSION……………………………………………………………….**28

**11.REFERENCES………………………………………………………………..**29-30

**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 abuse harassment or abuse 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?

**INTRODUCTION**

**Purpose**

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

**Scope**

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.

**LITERATURE SURVEY**

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

**Authors: Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown**

Abstract: We present a classifier to predict contextual polarity of subjective phrases in a sentence. Our approach features lexical scoring derived from the Dictionary of Affect in Language (DAL) and extended through WordNet, allowing us to automatically score the vast majority of words in our input avoiding the need for manual labeling. We augment lexical scoring with n-gram analysis to capture the effect of context. We combine DAL scores with syntactic constituents and then extract n-grams of constituents from all sentences. We also use the polarity of all syntactic constituents within the sentence as features. Our results show significant improvement over a majority class baseline as well as a more difficult baseline consisting of lexical n-grams.

**Robust sentiment detection on twitter from biased and noisy data**

**Authors: Luciano Barbosa and Junlan Feng**

Abstract: In this paper, we propose an approach to automatically detect sentiments on Twitter messages (tweets) that explores some characteristics of how tweets are written and meta-information of the words that compose these messages. Moreover, we leverage sources of noisy labels as our training data. These noisy labels were provided by a few sentiment detection websites over twitter data. In our experiments, we show that since our features are able to capture a more abstract representation of tweets, our solution is more effective than previous ones and also more robust regarding biased and noisy data, which is the kind of data provided by these sources.

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

**Authors: Michael Gamon.**

Abstract: We demonstrate that it is possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. We show that by using large feature vectors in combination with feature reduction, we can train linear support vector machines that achieve high classification accuracy on data that present classification challenges even for a human annotator. We also show that, surprisingly, the addition of deep linguistic analysis features to a set of surface level word n-gram features contributes consistently to classification accuracy in this domain.

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

**Authors: Gupta B, Negi M, Vishwakarma K, Rawat G & Badhani P**

Abstract: Twitter is a platform widely used by people to express their opinions and display sentiments on different occasions. Sentiment analysis is an approach to analyze data and retrieve sentiment that it embodies. Twitter sentiment analysis is an application of sentiment analysis on data from Twitter (tweets), in order to extract sentiments conveyed by the user. In the past decades, the research in this field has consistently grown. The reason behind this is the challenging format of the tweets which makes the processing difficult. The tweet format is very small which generates a whole new dimension of problems like use of slang, abbreviations etc. In this paper, we aim to review some papers regarding research in sentiment analysis on Twitter, describing the methodologies adopted and models applied, along with describing a generalized Python based approach.

**SYSTEM ANALYSIS**

**EXISTING SYSTEM**

Concept to analyse women safety using social networking messages and by applying machine learning algorithms on it. Now-a-days almost all peoples are using social networking sites to express their feelings and if any women feel unsafe in any area then she will express negative words in her post/tweets/messages and by analyzing those messages we can detect which area is more unsafe for women’s

**Drawback:**

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

**PROPOSED SYSTEM**

In proposed work we have used TWEEPY package from python to download tweets from twitter but every time INTERNET will not available to download tweets online so we downloaded MEETOO tweets on women safety and safe inside dataset folder. Application will read these tweets to detect women’s sentiments.

* Author using NLTK (natural language tool kit) to remove special symbols and stop words from tweets and to make them clean.
* Author using TEXTBLOB corpora package and dictionary to count positive, negative and neutral polarity and the tweets which has polarity value less than 0 will consider as negative as and greater than 0 and less than 0.5 will consider as neutral and polarity greater than 0.5 will consider as positive

**Advantages**

* Analysis of twitter texts collection also includes the name of people and name of women who stand up against abuse harassment and un ethical 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.

**SYSTEM REQUIREMENTS SPECIFICATIONS**

**SOFTWARE REQUIREMENTS**

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation.

The appropriation of requirements and implementation constraints gives the general overview of the project in regard to what the areas of strength and deficit are and how to tackle them.

* Python IDLE 3.7 version (or)
* Anaconda 3.7 (or)
* Jupiter (or)
* Google colab

**HARDWARE REQUIREMENTS**

Minimum hardware requirements are very dependent on the particular software being developed by a given Enthought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

Operating system : Windows, Linux

Processor : minimum intel i3

Ram : minimum 4 GB

Hard disk : minimum 250GB

**FUNCTIONAL AND NON-FUNCTIONAL REQUIREMENTS**

**FUNCTIONAL REQUIREMENTS**

1. Data Collection

2. Data Preprocessing

3. Training And Testing

4. Modeling

5. Predicting

**NON-FUNCTIONAL REQUIREMENTS**

NON-FUNCTIONAL REQUIREMENT (NFR) specifies the quality attribute of a software system. They judge the software system based on Responsiveness, Usability, Security, Portability and other non-functional standards that are critical to the success of the software system. Example of nonfunctional requirement, “how fast does the website load?” Failing to meet non-functional requirements can result in systems that fail to satisfy user needs. Non- functional Requirements allows you to impose constraints or restrictions on the design of the system across the various agile backlogs. Example, the site should load in 3 seconds when the number of simultaneous users are > 10000. Description of non-functional requirements is just as critical as a functional requirement.

* Usability requirement
* Serviceability requirement
* Manageability requirement
* Recoverability requirement
* Security requirement
* Data Integrity requirement
* Capacity requirement
* Availability requirement
* Scalability requirement
* Interoperability requirement
* Reliability requirement
* Maintainability requirement
* Regulatory requirement
* Environmental requirement

**SYSTEM STUDY**

**FEASIBILITY STUDY**

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

* ECONOMICAL FEASIBILITY
* TECHNICAL FEASIBILITY
* SOCIAL FEASIBILITY

**ECONOMICAL FEASIBILITY**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

**TECHNICAL FEASIBILITY**

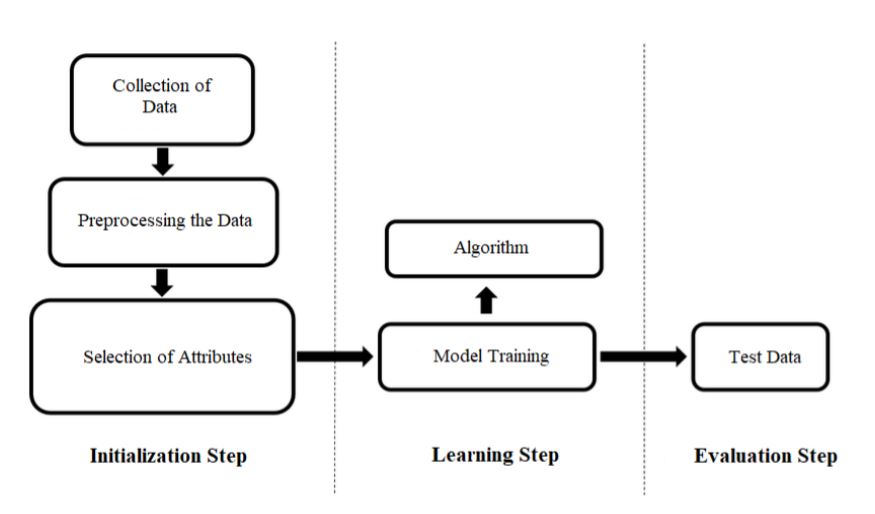
This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

**SOCIAL FEASIBILITY**

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user

**SYSTEM DESIGN**

**SYSTEM ARCHITECTURE:**



**DATA FLOW DIAGRAM**

* The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
* The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
* DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
* DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

Diagram

Description automatically generated

**UML DIAGRAMS**

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of objectoriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

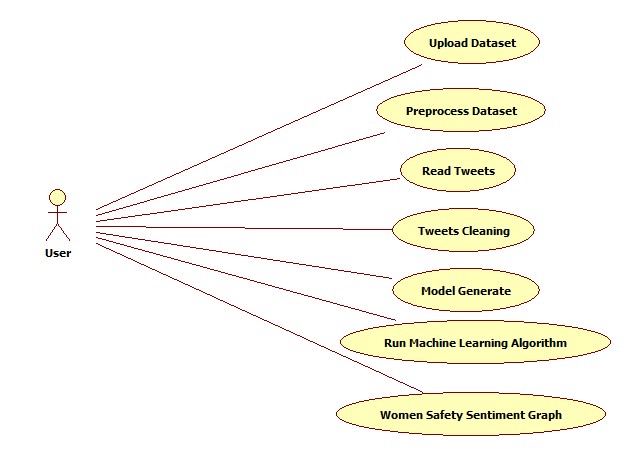
The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems. The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing Objects Oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

**GOALS:** The Primary goals in the design of the UML are as follows:

* Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
* Provide extendibility and specialization mechanisms to extend the core concepts.
* Be independent of particular programming languages and development process.
* Provide a formal basis for understanding the modeling language.
* Encourage the growth of OO tools market.
* Support higher level development concepts such as collaborations, frameworks, patterns and components.
* Integrate best practices.

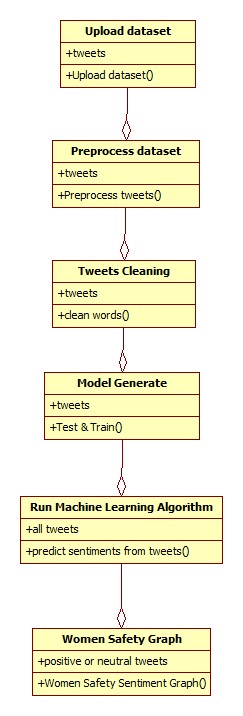
**USE CASE DIAGRAM:**

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



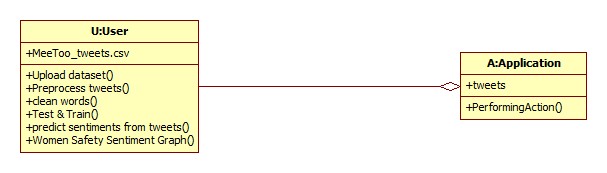
**Class diagram:**

The class diagram is used to refine the use case diagram and define a detailed design of the system. The class diagram classifies the actors defined in the use case diagram into a set of interrelated classes. The relationship or association between the classes can be either an "is-a" or "has-a" relationship. Each class in the class diagram may be capable of providing certain functionalities. These functionalities provided by the class are termed "methods" of the class. Apart from this, each class may have certain "attributes" that uniquely identify the class.



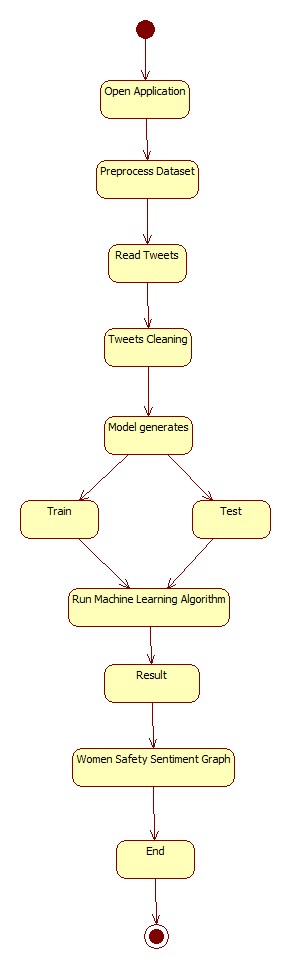
**Object diagram:**

The object diagram is a special kind of class diagram. An object is an instance of a class. This essentially means that an object represents the state of a class at a given point of time while the system is running. The object diagram captures the state of different classes in the system and their relationships or associations at a given point of time.



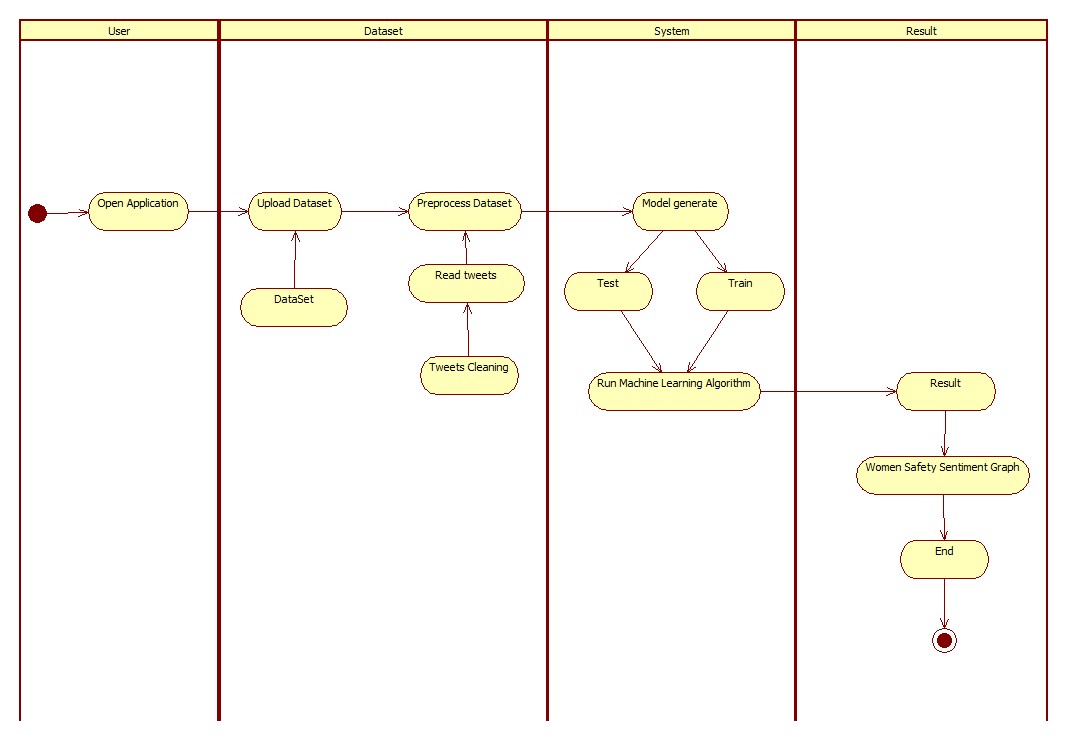
**State diagram:**

A state diagram, as the name suggests, represents the different states that objects in the system undergo during their life cycle. Objects in the system change states in response to events. In addition to this, a state diagram also captures the transition of the object's state from an initial state to a final state in response to events affecting the system.



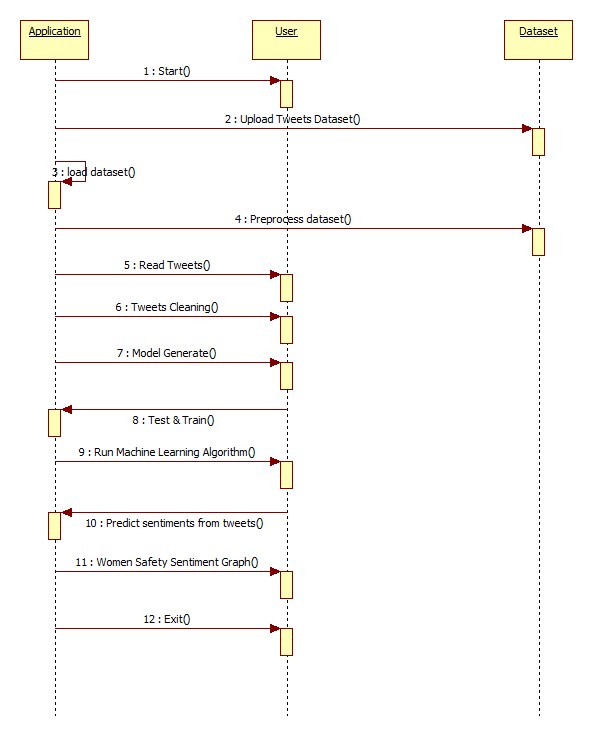
**Activity diagram:**

The process flows in the system are captured in the activity diagram. Similar to a state diagram, an activity diagram also consists of activities, actions, transitions, initial and final states, and guard conditions.



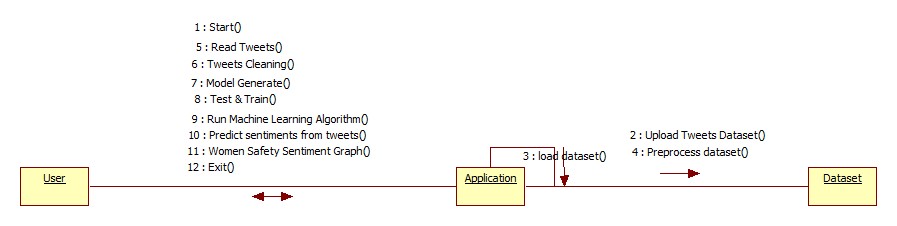
**Sequence diagram:**

A sequence diagram represents the interaction between different objects in the system. The important aspect of a sequence diagram is that it is time-ordered. This means that the exact sequence of the interactions between the objects is represented step by step. Different objects in the sequence diagram interact with each other by passing "messages".



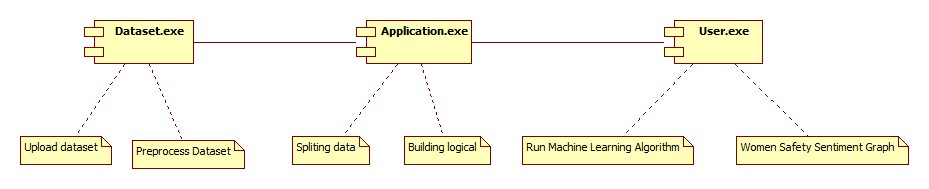
**Collaboration diagram:**

A collaboration diagram groups together the interactions between different objects. The interactions are listed as numbered interactions that help to trace the sequence of the interactions. The collaboration diagram helps to identify all the possible interactions that each object has with other objects.



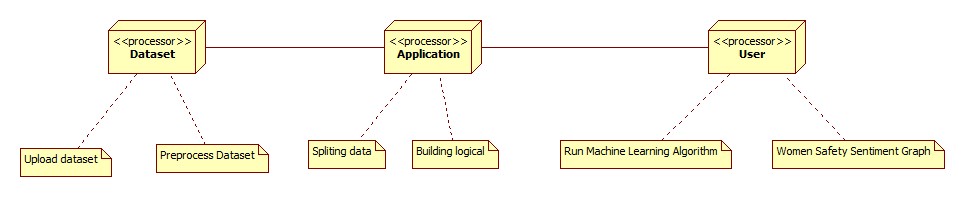
**Component diagram:**

The component diagram represents the high-level parts that make up the system. This diagram depicts, at a high level, what components form part of the system and how they are interrelated. A component diagram depicts the components culled after the system has undergone the development or construction phase.



**Deployment diagram:**

The deployment diagram captures the configuration of the runtime elements of the application. This diagram is by far most useful when a system is built and ready to be deployed.



**SYSTEM IMPLEMENTATION**

**MODULES**

* upload dataset: using this module we will upload dataset
* Dataset cleaning: using this module we will find out empty values in the dataset and replace with mean or 0 values.
* Train & Test Split: Using this module we will split dataset into two parts called and training and testing. All machine learning algorithms take 80% dataset to train classifier and 20% dataset is used to test classifier prediction accuracy.

**Implementation and Results**

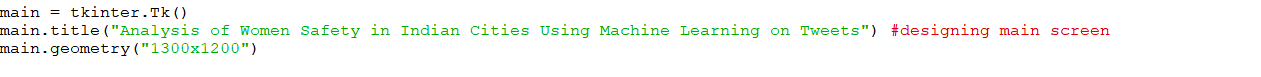
Step 1: import libraries

Graphical user interface, text, application, email

Description automatically generated

(where tkinter used for GUI(front-end),text blob -processing textual data, matplotlib - data visualization,pandas - data analysis and preprocessing,numpy - mathematical purpose,nltk - building python program( remove special symbols and stop words ))

Step 2:Defining main function and setting the title & size of tkinter



Step 3: Defining the global function

Background pattern

Description automatically generated with low confidence

Step 4: Upload data set

Graphical user interface, text

Description automatically generated

Step 5: Defining the read tweets function

Text

Description automatically generated with low confidence

Step 6: Defining tweet cleaning function

Text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Step 7: Build machine learning algorithm

Text

Description automatically generated

Step 8: Women safety graph

A picture containing text

Description automatically generated

Step 9: Defining the button size and configuration

Text, application

Description automatically generated

**SCREENSHOTS**

Now double click on ‘run.bat’ file to run project and to get below screen

Text

Description automatically generated

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

Graphical user interface

Description automatically generated

In above screen uploading MeeToo\_tweets.csv file and then click on ‘Open’ button to load dataset and to get below screen

Graphical user interface, application

Description automatically generated with medium confidence

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

Text

Description automatically generated

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

Text

Description automatically generated

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

Graphical user interface, application

Description automatically generated with medium confidence

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 Saftey Graph’ button to get below 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.

Chart, pie chart

Description automatically generated

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

**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.

**Future Enhancement:**

For the future enhancement, we can extend to apply these machine learning algorithms on different social media platforms like Facebook and Instagram also since in our project only twitter is considered. Present ideology which is proposed can be integrated with the twitter application interface to reach larger extent and apply sentimental analysis on millions of Tweet to provide more safety.

**REFERENCES**

Journals:

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.

Textbooks:

Programming Python, Mark Lutz

Head First Python, Paul Barry

Core Python Programming, R. Nageswara Rao

Learning with Python, Allen B. Downey

Websites:

<https://www.w3schools.com/python/>

<https://www.tutorialspoint.com/python/index.htm>

<https://www.javatpoint.com/python-tutorial>

<https://www.learnpython.org/>

https://www.pythontutorial.net/