Journal of Science and Technology
ISSN: 2456-5660 Volume 8, Issue 12 (Dec -2023)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2023.v8.i12.pp195-207*

Advancement in NLP with Decision Tree: The Impact of social media on Enhancing Women's Safety in Indian Cities

C. Gazala Akhtari[1], B. Vyhnavi[2], D. Deekshitha[2], Syed Sufiya Rana[2]

[1]Assistant Professor,[2]UG Students, Department of Cyber Security Engineering
[1,2]Malla Reddy Engineering College for Women, Maisammaguda, Dhulapally, Kompally,
Secunderabad-500100, Telangana, India.

To Cite this Article

C. Gazala Akhtari, B. Vyhnavi, D. Deekshitha, Syed Sufiya Rana "Advancement in NLP with Decision Tree: The Impact of social media on Enhancing Women's Safety in Indian Cities" *Journal of Science and Technology, Vol. 08, Issue 12 - Dec 2023, pp195-207*

Article Info

Received: 12-11-2023 Revised: 24 -11-2023 Accepted: 04-12-2023 Published: 14-12-2023

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. There have been several studies that have been conducted in cities across India and women report similar type of sexual harassment and passing off comments by other unknown people. The study that was conducted across most popular Metropolitan cities of India including Delhi, Mumbai, and Pune, it was shown that 60 % of the women feel unsafe while going out to work or while travelling in public transport. This work 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 work also focuses on how a sense of responsibility on part of Indian society can be developed the common Indian people so that they 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 they 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?

Keywords: Women's Safety, Social Media, Indian Cities, Metropolitan Cities.

1. INTRODUCTION

Staring and passing remarks are two examples of aggressive forms of harassment and violence, which are often seen as being a regular aspect of urban life [1]. Numerous studies have been undertaken in cities throughout India, and the results show that women report experiencing the same kinds of sexual harassment and passing remarks from other unidentified persons. According to research that was done in the most populated metropolises in India, including Delhi, Mumbai, and Pune [2], 60% of women report feeling dangerous when leaving the house for work or using public transportation. Women have the right to the city, which enables them to go anywhere they like, including to educational institutions and other destinations. However, because to the many unidentified Eyes body shaming and harassing these ladies, women feel frightened in settings like malls and shopping malls while traveling to their workplace [3]. Figure 1 shows the different types of abusive statics on women's in OSN environments. Here, the major cause of harassment of girls is safety or a lack of tangible repercussions in women's lives [4]. There are cases where girls are sexually harassed by their neighbours while they are walking to school. This is preferable than placing limits on women that society often places. With this method, we can automatically score the great majority of words in this input without the requirement for human labelling thanks to lexical scoring that is drawn from the Dictionary of Affect in Language (DAL) [5] and enhanced by WordNet. To account for the impact of context, they add n-gram analysis to the lexical scoring process. In order to extract n-grams of components from all sentences, they first integrate DAL scores with syntactic elements. Additionally, they employ the polarity of each syntactic element in the phrase as a feature [6]. A method to automatically identify feelings on Twitter communications (tweets) was suggested and analyses several writing characteristics of tweets as well as meta-information of the words that make up these messages. Additionally, they use sources of erratic labels as the training data. A couple emotion detection services over twitter data gave these noisy labels. Various trials demonstrate that machine learning technique [7] is superior to earlier ones and more resilient to skewed and noisy data, which is the kind of data offered by these sources, since these characteristics may capture a more abstract description of tweets.

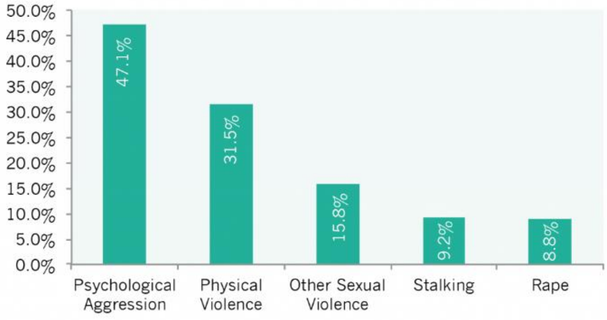


Figure 1. Women safety statistics in OSN.

The conventional methods [8] failed to predict the maximum safety analysis. As a result, the focus of this article is on the WSP-DT classifier. The Twitter dataset is initially considered to implement the entire system, which is then pre-processed to remove missing and unknown symbols. Then, NLTK was used to perform tokenization, lowercase conversion, stop word identification, stemming, and lemmatization on tweets. The text blob protocol is then developed to identify sentiments in pre-processed tweets, identifying positive, negative, and neutral polarities. The findings of a sentimental analysis can be put to use in a wide variety of contexts, such as determining people's perspectives on women; analysing public sentiments regarding government policies; determining people's attitudes toward a particular brand or the introduction of a new product; and so on. The information that was gathered from Twitter has been the subject of a substantial amount of study, which has been carried out in order to carry out the classification of tweets and assess the findings. This study also covers a number of other studies on machine learning, as well as research on how to carry out emotional analysis by making use of data from Twitter and applying it to a particular topic.

The algorithmic and model-based approaches to machine learning are the exclusive focus of this work. Some forms of violence and harassment, such as staring at women and making remarks, are commonplace even in urban settings, despite the fact that these behaviours, which are offensive and should not be tolerated, may be considered forms of violence and harassment. Numerous studies that have been carried out in India demonstrate that women have complained of being subjected to sexual harassment and other activities such as those mentioned above. Studies of this kind have also shown that the majority of women in populous metropolitan places such as Delhi, Pune, Chennai, and Mumbai have the feeling that they are in danger whenever they are surrounded by people they do not know.

2. LITERATURE SURVEY

these brief papers than in lengthier ones was looked in their study [9]. They make a number of observations on the difficulty of supervised learning for sentiment analysis in microblogs and are surprised to discover that identifying sentiment in microblogs is simpler than in blogs. As an example of a data mining technique [10], discussion of the study of the Twitter dataset using machine learning algorithms and sentiment analysis. A method for automatically categorizing Tweet emotions from the Twitter dataset is shown. Regarding a search keyword, these messages or tweets are categorized as favourable, negative, or neutral. This is highly helpful for businesses seeking feedback on their product brands or for consumers seeking third-party reviews of products before making a purchase. The sentiment of tweets [11] will be categorized using machine learning algorithms under remote supervision. Twitter tweets with emoticons and acronyms that serve as noisy labels make up the training data. They look at Twitter data's sentiment analysis. In [12] authors explored the unique field of doing sentiment analysis on people's perceptions of the best colleges in India. Spelling correction, which is neglected in previous research papers, was handled using a probabilistic model based on Bayes' theory in addition to other preprocessing steps including the expansion of net jargon and elimination of duplicate tweets [13]. The outcomes achieved by using the following machine learning techniques are contrasted in this study as well. Multilayer Perceptron is a model of an artificial neural network that combines Naive Bayes, SVM, and SVM. Additionally, a comparison of the four distinct SVM kernels—RBF, linear, polynomial, and sigmoid—has been made.

In [14], evaluation of various works on research into sentiment analysis on Twitter included a description. The sample of tweets was then subjected to sentiment analysis using a Python algorithm that had been trained using data mining. This allowed the tweets to be categorized based on the emotions they exhibited. The Sustainable Development Goals (SDGs) [15] were then utilized to categorize the tweets, and a textual analysis was conducted to determine the main environmental and public health issues that Twitter users are most concerned about. They accomplished this by using the qualitative analysis program NVivo Pro 12. In order to secure the safety of women nearby, in [16] authors concentrated on fostering duties among the general public in different Indian towns. Tweets sent with the Twitter app include text messages, audio, video, pictures, emoticons, and hashtags. This tweet content may be utilized to spread awareness among the public, enlightening them to the need to take stern action in the event that harassing tweets are sent out to women, and ultimately, punishing such individuals [17]. Twitter and Instagram, two platforms that support hashtags, may be used to disseminate messages throughout the world and encourage women to voice their opinions and sentiments. This will allow them to determine if they feel comfortable or not when they go for work, ride in a public vehicle, or are surrounded by ominous guys. This essay also focuses on how understanding [18] certain cultural norms might benefit average Indian people, emphasizing the need to protect the security of women around them. Tweets on Twitter, which mostly consist of images, messages, and remarks on the wellbeing of girls in Indian urban neighbourhoods, getting out of a public vehicle for work or a trip, the state of their psyche when mysterious men surround them, and whether or not these women have a strong sense of security.

In [19] authors explored how the use of machine learning methods in the investigative process may be used to study the pattern of crimes and criminal characteristics in India, such as rape, sexual assault, kidnapping, etc. The datasets collected from each state in the nation have been studied and worked on using the potent Python package pandas. In [20] authors scope includes an analysis of crime patterns as well as a focus on the fundamental causes of these patterns and the steps that should be done to avoid them going forward by using a decision tree's machine learning algorithm.

3. PROPOSED SYSTEM

3.1 Overview

In the work that was proposed, we downloaded tweets from Twitter using the TWEEPY package that is available for the Python programming language. However, every time we tried to download tweets online, the Internet was unavailable. As a result, we downloaded MEETOO tweets on women's safety and stored them in a dataset folder. This application will read these tweets in order to determine how ladies are feeling.

- Author is cleaning up tweets by using a programme called NLTK, which stands for natural language tool kit, in order to eliminate special symbols and stop words.
- The author uses the TEXTBLOB corpora package and dictionary to count positive, negative, and neutral polarity. Tweets with a polarity value of less than 0 are considered to have a negative polarity, while tweets with a polarity value of greater than 0 and less than 0.5 are considered to have a neutral polarity. Tweets with a polarity value of greater than 0.5 are considered to have a positive polarity.

every city is the prevalence of sexual harassment and assault. In addition, the harassing and violent content that can be found in online social networking sites can have a negative impact on the personal lives of women. As a result, it is essential to determine whether or not the OSN environment is safe for women. The conventional methods, on the other hand, were not successful in predicting the maximum safety analysis.

Figure 4.1 shows the proposed WPC-DT block diagram. Initially, dataset is collected using “TWEEPY” package, which download tweets from internet. The dataset mostly contains the “MEETOO” hashtag-based tweets. These tweets are specially focused on women safety issue. Then, the dataset is pre-processed using NLTK. Here, NLTK is used to identify stop words, and remove special symbols from tweet dataset. The NLTK also eliminates unknown characters, symbols, special letters from dataset. The empty samples are replaced by zeros, which resulted in pre-processed and normalized data.

Then, Textblob is used to count the positive, negative, and neutral polarity tweets. Tweets with polarity values less than 0 are considered negative, tweets with polarity values greater than 0 are considered neutral, and tweets with polarity greater than 0.5 are considered positive. Further, TF-IDF method is used to extract the data specific features. In addition, DT classifier trained with TF-IDF features. Finally, The DT classifier predicts the tweet status as “Genuine tweet” or “Fake tweet” by using sentiment analysis.

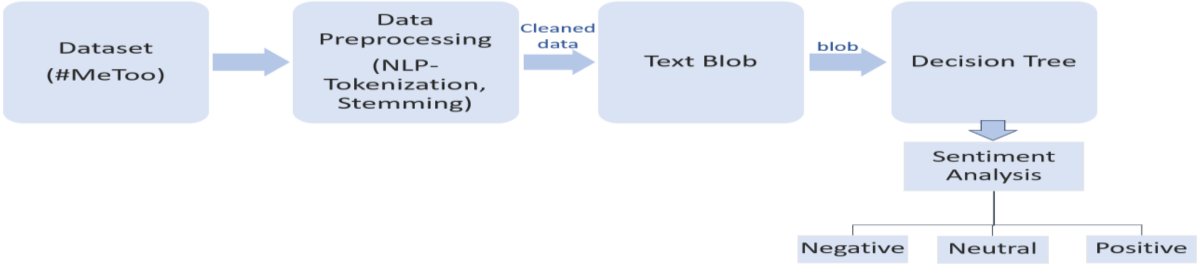


Figure 2. Block diagram of proposed system.

3.2 Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Step 1: Load the Hotel Review Dataset: In this step, load the dataset into your data analysis environment. This dataset typically includes text reviews (the input) and corresponding labels (the output), which can be sentiments like "positive" or "negative." The goal is to train a model to predict these labels based on the text.

Step 2: Data Cleaning: Data cleaning involves several sub-steps: Removing Special Characters and Punctuation: Special characters (e.g., @, $, %) and punctuation (e.g., !, ?, .) are often irrelevant to sentiment analysis and can be removed to focus on the actual text content. Handling Irrelevant Information: Sometimes, there might be metadata or other information in the text data that's not relevant to the analysis. You should remove such information to concentrate on the review text itself.

Step 3: Tokenization:

- Tokenization is the process of splitting the text into smaller units, such as words or phrases (tokens). This step is crucial because it breaks down the text into manageable pieces for further analysis.
- For example, the sentence "I love this hotel" would be tokenized into ["I", "love", "this", "hotel"].

Step 4: Convert Text to Lowercase:

- Converting all text to lowercase ensures consistency in your text data. It prevents the model from treating "good" and "Good" as two different words, which could lead to incorrect feature extraction and modeling.
- For example, "Good" and "good" should be treated as the same word.

Step 5: Remove Stop Words:

- Stop words are common words in a language that often don't carry significant meaning and can be safely removed to reduce noise in the data. Examples include "the," "is," "and," "in," etc.
- Removing stop words can help improve the efficiency of the model and reduce the dimensionality of the data without losing much valuable information.

Step 6: Apply Stemming or Lemmatization:

- Stemming and lemmatization are techniques used to reduce words to their root form, which helps in standardizing words and improving feature extraction.

- Stemming: It involves removing suffixes from words to obtain the word stem. For example, "jumping" becomes "jump," "flies" becomes "fli," etc. Stemming is more aggressive but may result in non-words.
- Lemmatization: It is a more advanced technique that reduces words to their base or dictionary form (lemma). For example, "better" becomes "good," "running" becomes "run," etc. Lemmatization is more accurate but computationally expensive.
- The choice between stemming and lemmatization depends on your specific NLP task and dataset.

3.3 Dataset Splitting

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So, we always try to make a machine learning model which performs well with the training set and with the test dataset.

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

3.4 TF-IDF Feature Extraction

TF-IDF which stands for Term Frequency – Inverse Document Frequency. It is one of the most important techniques used for information retrieval to represent how important a specific word or phrase is to a given document. Let's take an example, we have a string or Bag of Words (BOW) and we have to extract information from it, then we can use this approach.

Figure 4.2 shows the TF-IDF feature extraction block diagram. The tf-idf value increases in proportion to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps to adjust with respect to the fact that some words appear more frequently in general. TF-IDF use two statistical methods, first is Term Frequency and the other is Inverse Document Frequency. Term frequency refers to the total number of times a given term t appears in the document doc against (per) the total number of all words in the document and the inverse document frequency measure of how much information the word provides. It measures the weight of a given word in the entire document. IDF show how common or rare a given word is across all documents. TF-IDF can be computed as $tf * idf$.

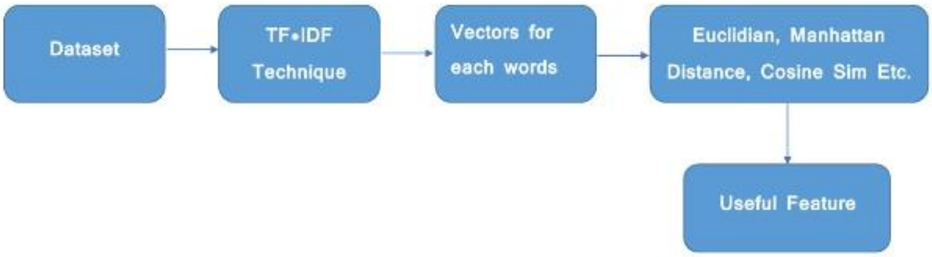


Figure. 3: TF-IDF block diagram.

TF-IDF do not convert directly raw data into useful features. Firstly, it converts raw strings or dataset into vectors and each word has its own vector. Then we'll use a particular technique for retrieving the feature like Cosine Similarity which works on vectors, etc.

Terminology

t — term (word)

d — document (set of words)

N — count of corpus

corpus — the total document set

Step 1: Term Frequency (TF): Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, "Data Science is awesome!" A simple way to start out is by eliminating documents that do not contain all three words "Data" is", "Science", and "awesome", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency. The weight of a term that occurs in a document is simply proportional to the term frequency.

$$tf(t,d) \ = \ count \ of \ t \ in \ d \ / \ number \ of \ words \ in \ d$$

Step 2: Document Frequency: This measures the importance of document in whole set of corpora, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N. In other words, DF is the number of documents in which the word is present. We consider one occurrence if the term consists in the document at least once, we do not need to know the number of times the term is present.

$$df(t) \ = \ occurrence \ of \ t \ in \ documents$$

Step 3: Inverse Document Frequency (IDF): While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus, we need to weigh down the frequent terms while scale up the rare ones, by computing IDF, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. The IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as "is" is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

$$idf(t) \ = \ N/df$$

Now there are few other problems with the IDF, in case of a large corpus, say 100,000,000, the IDF value explodes, to avoid the effect we take the log of idf . During the query time, when a word which is not in vocab occurs, the df will be 0. As we cannot divide by 0, we smoothen the value by adding 1 to the denominator.

$$idf(t) \ = \ log(N/(df \ + \ 1))$$

The TF-IDF now is at the right measure to evaluate how important a word is to a document in a collection or corpus. Here are many different variations of TF-IDF but for now let us concentrate on this basic version.

$$tf - idf(t,d) = tf(t,d) * log(N/(df + 1))$$

4. RESULTS AND DISCUSSION

4.1 Implementation description

This is a Python script that creates a graphical user interface (GUI) using the tkinter library for a machine learning application. The application's purpose is to analyze tweets related tothe safety of women in Indian cities. It performs the following tasks:

- GUI Initialization:
  It starts by creating a tkinter window named main.
  The title of the window is set to "Machine Learning Application: The role of social media in Promoting the Safety of Women in Indian Cities".
  The window's dimensions are set to 1300x1200 pixels.
- Importing Libraries:
  The script imports necessary libraries including tkinter for GUI, TextBlob for sentiment analysis, matplotlib for plotting graphs, numpy and pandas for data manipulation, and others.
- Function Definitions:
  tweetCleaning: Cleans a given tweet by removing punctuation, non-alphabetic characters, stopwords, and short words.
- upload: Allows the user to upload a dataset containing tweets.
  read: Reads and displays the tweets from the uploaded dataset.
  clean: Cleans the tweets using the tweetCleaning function.
  machineLearning: Applies sentiment analysis using TextBlob and categorizes tweets into positive, negative, or neutral sentiments.
  graph: Generates a pie chart to visualize the distribution of sentiments.
- Global Variables:
  filename: Holds the path of the uploaded file.
  tweets_list: Stores the original tweets.
  clean_list: Stores the cleaned tweets.
  pos, neu, neg: Counters for positive, neutral, and negative sentiments.

- Button Definitions:
  Buttons are created for actions like uploading a dataset, reading tweets, cleaning tweets, applying machine learning, and generating a graph. Each button is associated with a specific function.
- Label and Text Box:
  Labels and a text box are added to the GUI for displaying information and results.
- Execution:
  The GUI window is configured with background color and launched with main.mainloop().
- Event Handling:
  When a button is clicked, it triggers the associated function, performing the desired action.
- Displaying Results:

The GUI displays various information like the uploaded file's path, total number of tweets, cleaned tweets, and sentiment analysis results.

- Graph Plotting:
  The graph function displays a pie chart showing the distribution of positive, negative, and neutral sentiments.

This script utilizes various libraries and functions for specific tasks, such as TextBlob for sentiment analysis and tkinter for GUI. The application is designed to be user-friendly, allowing the user to interact with it through the GUI.

4.2 Dataset description

The dataset contains following columns

- Text: This column contains the actual content of the tweets posted by users. It includes messages, statements, or comments related to the #MeToo movement. The text may include user mentions, hashtags, links, and any other information the user has posted.
- Id: This column likely serves as a unique identifier for each tweet. It provides a way to differentiate one tweet from another and is typically assigned by the platform (Twitter).
- Length: This column indicates the length of each tweet in terms of the number of characters. It provides quantitative information about the size of the tweet's content.
- Created_at: This column records the timestamp indicating when each tweet was posted. It provides information about the date and time of tweet creation.
- Source: The "Source" column specifies the application or platform used by the user to post the tweet. This information can be valuable for understanding user behavior and preferences regarding tweet creation.
- Favorite_count: This column represents the number of times a particular tweet was marked as a favorite or liked by other users. It indicates the level of engagement and popularity of each tweet.
- Retweet_count: The "Retweet_count" column indicates how many times a specific tweet was retweeted by other users. It provides insights into the extent to which the tweet's content was shared and disseminated within the Twitter community.
- Lang: The "Lang" column specifies the language in which each tweet was posted. For instance, 'en' indicates English.

4.3 Results and Description

Figure 10.1 portrays the graphical user interface (GUI) design of this work's proposed system. This interface serves as the visual gateway through which users engage with the system's functionalities. Within this graphical environment, various interactive elements are likely presented, such as buttons, input fields, and display areas. Users can utilize these elements to perform a range of actions, including uploading datasets, initiating sentiment analysis, and reviewing the results. The GUI is a critical aspect of this work as it acts as a bridge between users and the system's underlying processes. It provides an intuitive means for users to navigate, input data, trigger operations, and visualize the outcomes of sentiment analysis, enhancing the overall user experience and facilitating efficient interaction with the system.

Figure 10.2 likely illustrates the process of uploading and reading datasets within the system. This image could depict a screen or dialog where users can select and import datasets into the system. It signifies the initial step in data analysis, allowing users to provide the raw data that will be subsequently processed and analyzed by the system.

Figure 10.3 showcases a snapshot of the dataset post-preprocessing. The preprocessing phase involves a series of steps like data cleaning, tokenization, and removal of irrelevant information. This figure may present a view of the dataset in a more organized and structured format, prepared for analysis. It provides users with an insight into how the data has been refined for subsequent tasks.

Figure 10.4 appears to exhibit the outcomes of sentiment analysis conducted on a test tweet. It possibly displays information such as the system's prediction regarding the sentiment of the tweet (whether it's negative, neutral, or positive) and a polarity score, typically ranging from 0 to 1, indicating the strength or intensity of the sentiment expressed in the tweet. This visual aids users in understanding how the system categorizes and scores sentiment within individual tweets.

Figure 10.5 likely provides a visual representation, such as a bar or pie chart, demonstrating the distribution of sentiment within the analyzed dataset. The percentages mentioned in your description (e.g., 74.6% negative, 22.3% neutral, 3.1% positive) indicate the proportion of tweets falling into each sentiment category. This graphical representation offers an at-a-glance overview of the sentiment landscape within the dataset, helping users assess sentiment patterns and trends.

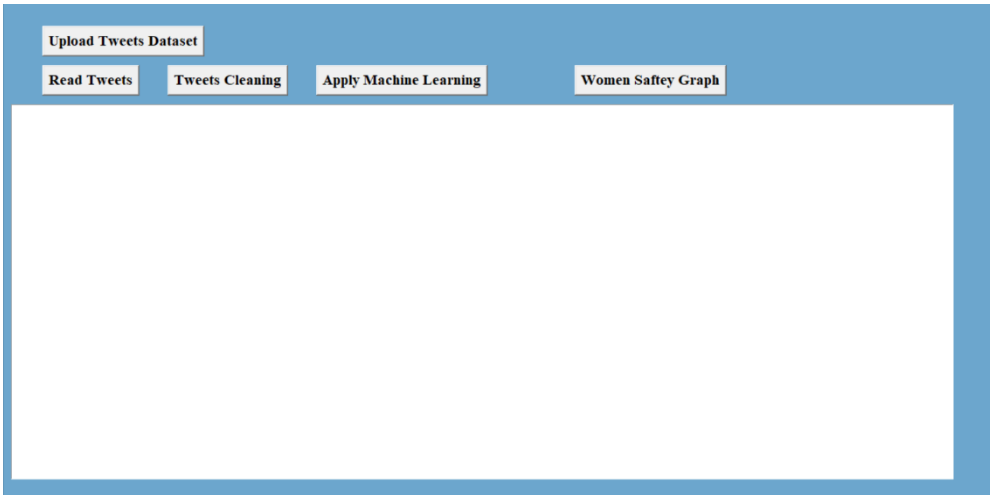
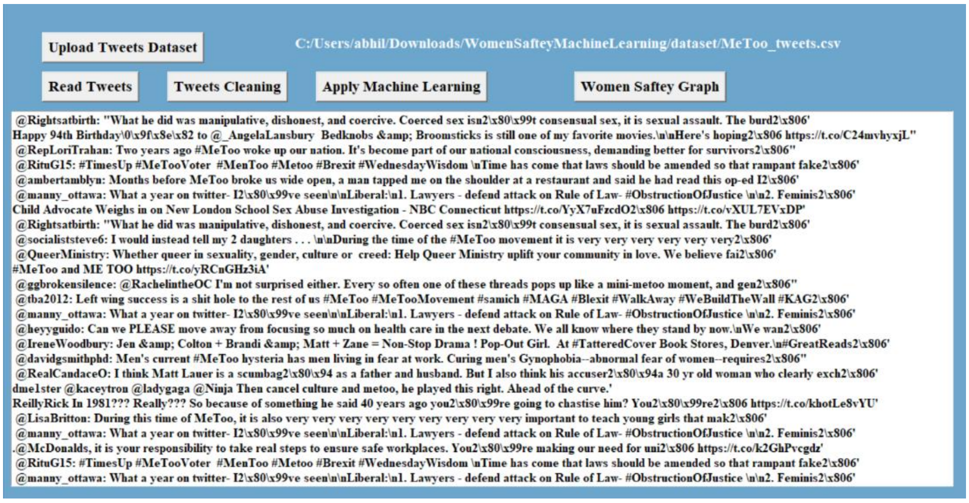


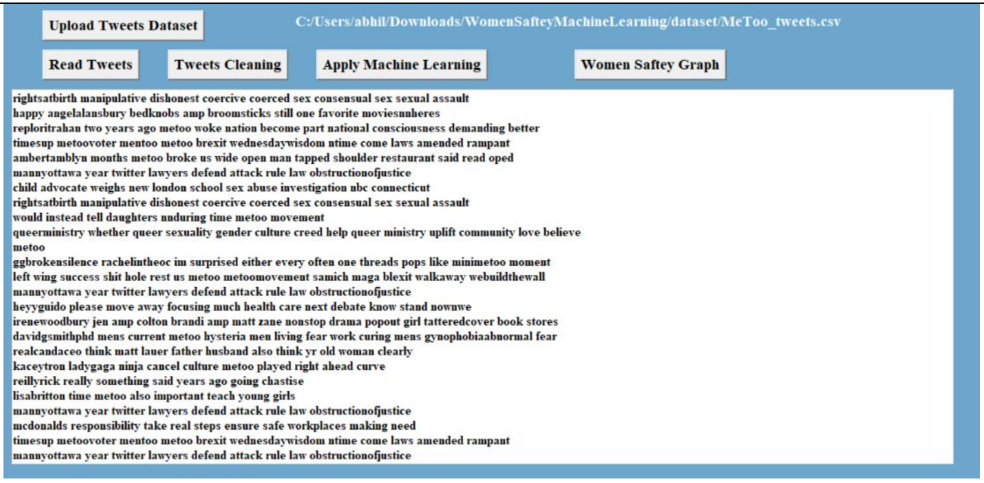
Figure 4: User interface of proposed system.



```
C:/Users/abhil/Downloads/WomenSafteyMachineLearning/dataset/MeToo_tweets.csv
```

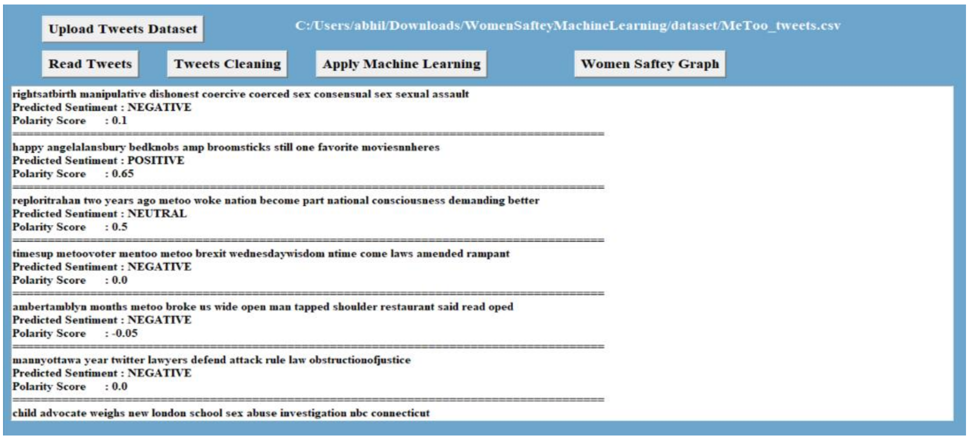
```
 @Rightsatbirth: "What he did was manipulative, dishonest, and coercive. Coerced sex isn2\x80\x99t consensual sex, it is sexual assault. The burd2\x806'
Happy 94th Birthday\0\x9f\x8e\x82 to @_AngelaLansbury  Bedknobs &amp; Broomsticks is still one of my favorite movies.\n\nHere's hoping2\x806 https://t.co/C24mvhyxjL"
 @RepLoriTrahan: Two years ago #MeToo woke up our nation. It's become part of our national consciousness, demanding better for survivors2\x806"
 @RituG15: #TimesUp #MeTooVoter  #MenToo #Metoo #Brexit #WednesdayWisdom \nTime has come that laws should be amended so that rampant fake2\x806'
 @ambertamblyn: Months before MeToo broke us wide open, a man tapped me on the shoulder at a restaurant and said he had read this op-ed I2\x806'
 @manny_ottawa: What a year on twitter- I2\x80\x99ve seen\n\nLiberal:\n1. Lawyers - defend attack on Rule of Law- #ObstructionOfJustice \n\n2. Feminis2\x806'
Child Advocate Weighs in on New London School Sex Abuse Investigation - NBC Connecticut https://t.co/YyX7uFzcdO2\x806 https://t.co/vXUL7EVxDP'
 @Rightsatbirth: "What he did was manipulative, dishonest, and coercive. Coerced sex isn2\x80\x99t consensual sex, it is sexual assault. The burd2\x806'
 @socialiststeve6: I would instead tell my 2 daughters . . . \n\nDuring the time of the #MeToo movement it is very very very very very very2\x806'
 @QueerMinistry: Whether queer in sexuality, gender, culture or  creed: Help Queer Ministry uplift your community in love. We believe fai2\x806'
#MeToo and ME TOO https://t.co/yRCnGHz3iA'
 @ggbrokensilence: @RachelintheOC I'm not surprised either. Every so often one of these threads pops up like a mini-metoo moment, and gen2\x806"
 @tba2012: Left wing success is a shit hole to the rest of us #MeToo #MeTooMovement #samich #MAGA #Blexit #WalkAway #WeBuildTheWall #KAG2\x806'
 @manny_ottawa: What a year on twitter- I2\x80\x99ve seen\n\nLiberal:\n1. Lawyers - defend attack on Rule of Law- #ObstructionOfJustice \n\n2. Feminis2\x806'
 @heyyguido: Can we PLEASE move away from focusing so much on health care in the next debate. We all know where they stand by now.\nWe wan2\x806'
 @IreneWoodbury: Jen &amp; Colton + Brandi &amp; Matt + Zane = Non-Stop Drama ! Pop-Out Girl.  At #TatteredCover Book Stores, Denver.\n#GreatReads2\x806'
 @davidgsmithphd: Men's current #MeToo hysteria has men living in fear at work. Curing men's Gynophobia--abnormal fear of women--requires2\x806"
 @RealCandaceO: I think Matt Lauer is a scumbag2\x80\x94 as a father and husband. But I also think his accuser2\x80\x94a 30 yr old woman who clearly exch2\x806'
dme1ster @kaceytron @ladygaga @Ninja Then cancel culture and metoo, he played this right. Ahead of the curve.'
ReillyRick In 1981??? Really??? So because of something he said 40 years ago you2\x80\x99re going to chastise him? You2\x80\x99re2\x806 https://t.co/khotLe8vYU'
 @LisaBritton: During this time of MeToo, it is also very very very very very very very very very important to teach young girls that mak2\x806'
 @manny_ottawa: What a year on twitter- I2\x80\x99ve seen\n\nLiberal:\n1. Lawyers - defend attack on Rule of Law- #ObstructionOfJustice \n\n2. Feminis2\x806'
.@McDonalds, it is your responsibility to take real steps to ensure safe workplaces. You2\x80\x99re making our need for uni2\x806 https://t.co/k2GhPvcgdz'
 @RituG15: #TimesUp #MeTooVoter  #MenToo #Metoo #Brexit #WednesdayWisdom \nTime has come that laws should be amended so that rampant fake2\x806'
 @manny_ottawa: What a year on twitter- I2\x80\x99ve seen\n\nLiberal:\n1. Lawyers - defend attack on Rule of Law- #ObstructionOfJustice \n\n2. Feminis2\x806'
```

Figure 5: Dataset upload and reading.



```
rightsatbirth manipulative dishonest coercive coerced sex consensual sex sexual assault
happy angelalansbury bedknobs amp broomsticks still one favorite moviesnnheres
reploritrahan two years ago metoo woke nation become part national consciousness demanding better
timesup metoovoter mentoo metoo brexit wednesdaywisdom ntime come laws amended rampant
ambertamblyn months metoo broke us wide open man tapped shoulder restaurant said read oped
mannyottawa year twitter lawyers defend attack rule law obstructionofjustice
child advocate weighs new london school sex abuse investigation nbc connecticut
rightsatbirth manipulative dishonest coercive coerced sex consensual sex sexual assault
would instead tell daughters nnduring time metoo movement
queerministry whether queer sexuality gender culture creed help queer ministry uplift community love believe
metoo
ggbrokensilence rachelintheoc im surprised either every often one threads pops like minimetoo moment
left wing success shit hole rest us metoo metoomovement samich maga blexit walkaway webuildthewall
mannyottawa year twitter lawyers defend attack rule law obstructionofjustice
heyyguido please move away focusing much health care next debate know stand nownwe
irenewoodbury jen amp colton brandi amp matt zane nonstop drama popout girl tatteredcover book stores
davidgsmithphd mens current metoo hysteria men living fear work curing mens gynophobiaabnormal fear
realcandaceo think matt lauer father husband also think yr old woman clearly
kaceytron ladygaga ninja cancel culture metoo played right ahead curve
reillyrick really something said years ago going chastise
lisabritton time metoo also important teach young girls
mannyottawa year twitter lawyers defend attack rule law obstructionofjustice
mcdonalds responsibility take real steps ensure safe workplaces making need
timesup metoovoter mentoo metoo brexit wednesdaywisdom ntime come laws amended rampant
mannyottawa year twitter lawyers defend attack rule law obstructionofjustice
```

Figure 6: Dataset after preprocessing.



```
rightsatbirth manipulative dishonest coercive coerced sex consensual sex sexual assault
Predicted Sentiment : NEGATIVE
Polarity Score      : 0.1
================================================================================
happy angelalansbury bedknobs amp broomsticks still one favorite moviesnnheres
Predicted Sentiment : POSITIVE
Polarity Score      : 0.65
================================================================================
reploritrahan two years ago metoo woke nation become part national consciousness demanding better
Predicted Sentiment : NEUTRAL
Polarity Score      : 0.5
================================================================================
timesup metoovoter mentoo metoo brexit wednesdaywisdom ntime come laws amended rampant
Predicted Sentiment : NEGATIVE
Polarity Score      : 0.0
================================================================================
ambertamblyn months metoo broke us wide open man tapped shoulder restaurant said read oped
Predicted Sentiment : NEGATIVE
Polarity Score      : -0.05
================================================================================
mannyottawa year twitter lawyers defend attack rule law obstructionofjustice
Predicted Sentiment : NEGATIVE
Polarity Score      : 0.0
================================================================================
child advocate weighs new london school sex abuse investigation nbc connecticut
```

Figure 7: Prediction results from test tweet.

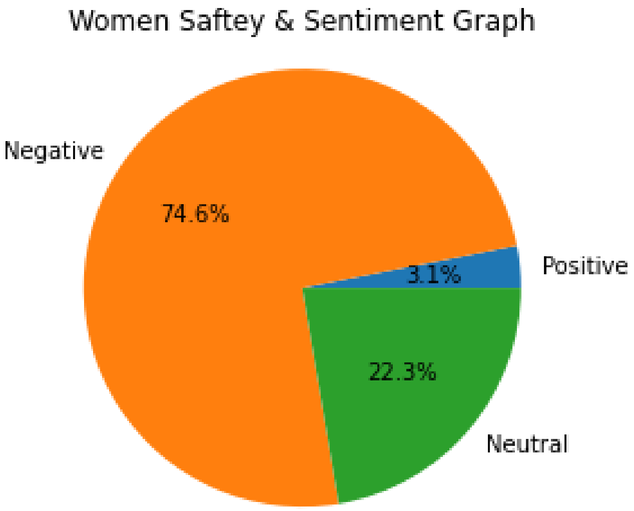


Figure 8: Sentiment graph performance measurement.

5. Conclusion

In conclusion, the proposed WPC-DT (Women's Safety Tweet Prediction Classifier) system, outlines a comprehensive approach to addressing the issue of identifying genuine and fake tweets related to women's safety, particularly those using the "MEETOO" hashtag. The process begins with data collection through the "TWEEPY" package, which downloads tweets from the internet, focusing on the specified hashtag. Subsequently, the dataset undergoes rigorous pre-processing using NLTK (Natural Language Toolkit). NLTK plays a crucial role in cleaning and normalizing the data, removing stop words, special symbols, unknown characters, and special letters, while handling empty samples by replacing them with zeros. This meticulous pre-processing results in a refined and normalized dataset.

Following pre-processing, Textblob is employed to analyze the sentiment of the tweets, categorizing them into positive, negative, or neutral based on their polarity scores. This sentiment analysis helps distinguish between tweets that express negative sentiments, neutral sentiments, and those with strong positive sentiments, with a polarity threshold of 0.5 defining "positive." Furthermore, the TF-IDF (Term Frequency-Inverse Document Frequency) method is applied to extract specific features from the data, enhancing the system's ability to identify unique characteristics of tweets related to women's safety. These TF-IDF features are then utilized to train a Decision Tree (DT) classifier.

In the final step, the trained DT classifier is employed to predict the status of tweets as either "Genuine" or "Fake" by leveraging the results of the sentiment analysis. This integration of sentiment analysis and machine learning techniques enhances the accuracy and reliability of tweet classification, contributing to the broader discourse on women's safety.

REFERENCES

[1]. Gamon and 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.

[2]. 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.

[3]. 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.

[4]. Bermingham, Adam, and A. F. Smeaton. “Classifying sentiment in microblogs: is brevity an advantage?”, Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[5]. V. Sahayak, V. Shete, and A. Pathan (2015). “Sentiment analysis on twitter data. International Journal of Innovative Research in Advanced Engineering (IJIRAE)”, 2(1), 178-183.

[6]. N. Mamgain, E. Mehta, A. Mittal and G. Bhatt, “Sentiment analysis of top colleges in India using Twitter data”, 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016, pp. 525-530, doi: 10.1109/ICCTICT.2016.7514636.

[7]. B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, and P. Badhani (2017). “Study of Twitter sentiment analysis using machine learning algorithms on Python”. International Journal of Computer Applications, 165(9), 0975-8887.

[8]. Reyes-Menendez, J. R. Saura, and C. Alvarez Alons. “Understanding# World Environment Day user opinions in Twitter: A topic-based sentiment analysis approach”. International journal of environmental research and public health. 2018 Nov;15(11):2537.

[9]. D. Kumar and S. Aggarwal. “Analysis of Women Safety in Indian Cities Using Machine Learning on Tweets”, 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 159-162, doi: 10.1109/AICAI.2019.8701247.

[10]. Vikram Chandra and Rampur Srinath. “Analysis of Women Safety using Machine Learning on Tweets”, (IRJET) 2020.

[11]. F. Bravo-Marquez, B. Pfahringer, S. Mohammad and E. Frank, “Affective Tweets: a Weka Package for Analysing effect in Tweets”, Journal of Machine Learning Research, vol. 20, no. 92, pp. 1-6, 2020.

[12]. K. Abdul Sattar, Q. Obeidat and M. Akure. “Towards harnessing based learning algorithms for tweets sentiment analysis international conference of innovation and intelligence for informatics Computing and technology 2020”.

[13]. K. R. Teja, K. A. Kumar, G. S. Praveen and D. N. Harini. “Analysis of Crimes Against Women in India Using Machine Learning Techniques”, In Communication Software and Networks 2021 (pp. 499-510). Springer, Singapore.

[14]. Srinivasan, S., P. Muthu Kannan, and R. Kumar. "A Machine Learning Approach to Design and Develop a BEACON Device for Women’s Safety." Recent Advances in Internet of Things and Machine Learning. Springer, Cham, 2022. 111-115.

[15]. Bonny, Afrin Jaman, et al. "Sentiment Analysis of User-Generated Reviews of Women Safety Mobile Applications." 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). IEEE, 2022.

[16]. Ashok, K., et al. "A Survey on Design and Application Approaches in Women-Safety Systems." 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS). Vol. 1. IEEE, 2022.

[17]. Tran, Martino, et al. "Monitoring the well-being of vulnerable transit riders using machine learning based sentiment analysis and social media: Lessons from COVID-19." Environment and Planning B: Urban Analytics and City Science (2022): 23998083221104489.

[18]. Zhong, Yongqi, et al. "Use of machine learning to estimate the per-protocol effect of low-dose aspirin on pregnancy outcomes: a secondary analysis of a randomized clinical trial." JAMA network open 5.3 (2022): e2143414-e2143414.

[19]. Patel, Bansi, and Manmitsinh C. Zala. "Crime Against Women Analysis & Prediction in India Using Supervised Regression." 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). IEEE, 2022.

[20]. Islam, Md M., et al. "Risk factors identification and prediction of anemia among women in Bangladesh using machine learning techniques." Current Women's Health Reviews 18.1 (2022): 118-133