MACHINE LEARNING APPLICATION: THE ROLE OF SOCIAL MEDIA IN PROMOTING THE SAFETY OF WOMEN IN INDIAN CITIES

B. Durga Bhavani[1], S. Vaishnavi[2], T. Akshara[2], S. Vaishnavi[2], V. Harini[2]

[1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

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 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 safety, social media, machine learning.

1. INTRODUCTION

There are certain types of harassment and Violence that are very aggressive including staring and passing comments and these unacceptable practices are usually seen as a normal part of the urban life. 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. 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

Safety or lack of concrete consequences in the life of women is the main reason of harassment of girls. There are instances when the harassment of girls was done by their neighbours while they were on the way to school or there was a lack of safety that created a sense of fear in the minds of small girls who throughout their lifetime suffer due to that one instance that happened in their lives where they were forced to do something unacceptable or was sexually harassed by one of their own neighbors or any other unknown person. Safest cities approach women safety from a perspective of women rights to the affect the city without fear of violence or sexual harassment. Rather than imposing restrictions on women that society usually imposes it is the duty of society to imprecise the need of protection of women and also recognizes that women and girls also have a right same as men have to be safe in the City.

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.

Gamon and Michael et. al [1] demonstrate that it is possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. They show that by using large feature vectors in combination with feature reduction, they can train linear support vector machines that achieve high classification accuracy on data that present classification challenges even for a human annotator. They 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. Agarwal et. al [2] presented a classifier to predict contextual polarity of subjective phrases in a sentence. This 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 this input avoiding the need for manual labelling. They augment lexical scoring with n-gram analysis to capture the effect of context. They combine DAL scores with syntactic constituents and then extract n-grams of constituents from all sentences. They also use the polarity of all syntactic constituents within the sentence as features.

Barbosa et. al [3] proposed 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, they leverage sources of noisy labels as this training data. These noisy labels were provided by a few sentiment detection websites over twitter data. In these experiments, they show that since these 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.

2. LITERATURE SURVEY

Bermingham et. al [4] examined the hypothesis that it is easier to classify the sentiment in these short form documents than in longer form documents. Surprisingly, they find classifying sentiment in microblogs easier than in blogs and make a number of observations pertaining to the challenge of supervised learning for sentiment analysis in microblogs.

Sahayak et. al [5] discusses the existing analysis of twitter dataset with data mining approach such as use of Sentiment analysis algorithm using machine learning algorithms. An approach is introduced that automatically classifies the sentiments of Tweets taken from Twitter dataset. These messages or tweets are classified as positive, negative, or neutral with respect to a query term. This is very useful for the companies who want to know the feedback about their product brands or the customers who want to search the opinion from others about product before purchase. They will use machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision. The training data consists of Twitter messages with emoticons, acronyms which are used as noisy labels. They examine sentiment analysis on Twitter data.

Mamgain et. al [6] to dive into the novel domain of performing sentiment analysis of people's opinions regarding top colleges in India. Besides taking additional preprocessing measures like the expansion of net lingo and removal of duplicate tweets, a probabilistic model based on Bayes' theorem was used for spelling correction, which is overlooked in other research studies. This paper also highlights a comparison between the results obtained by exploiting the following machine learning algorithms: Naïve Bayes and Support Vector Machine and an Artificial Neural Network model: Multilayer Perceptron. Furthermore, a contrast has been presented between four different kernels of SVM: RBF, linear, polynomial, and sigmoid.

Gupta et. al [7] aimed 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.

Reyes-Menendez et. al [8] identified the social, economic, environmental, and cultural factors related to the sustainable care of both environment and public health that most concern Twitter users with 336 million active users as of 2018, Twitter is a social network that is increasingly used in research to get information and to understand public opinion as exemplified by Twitter users. In order to identify the factors related to the sustainable care of environment and public health, they have downloaded n = 5873 tweets that used the hashtag #WorldEnvironmentDay on the respective day. As the next step, sentiment analysis with an algorithm developed in Python and trained with data mining was applied to the sample of tweets to group them according to the expressed feelings. Thereafter, a textual analysis was used to group the tweets according to the Sustainable Development Goals (SDGs), identifying the key factors about environment and public health that most concern Twitter users. To this end, they used the qualitative analysis software NVivo Pro 12.

Kumar and Aggarwal et. al [9] 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 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?

3. PROPOSED SYSTEM

Fig. 3.1 shows the proposed block diagram analysis of women safety. 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 natural language toolkit (NLTK). Here, NLTK is used to remove special symbols and stop words 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, Text blob is used to count the positive, negative, and neutral polarity tweets it 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.

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.

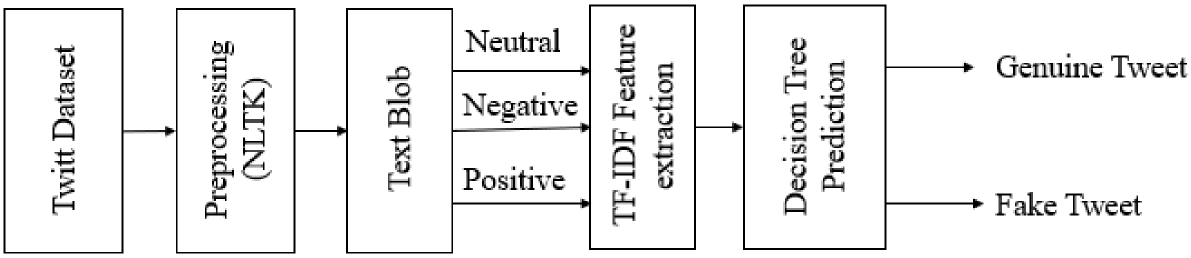


Fig. 1: Block diagram of proposed system.

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.

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

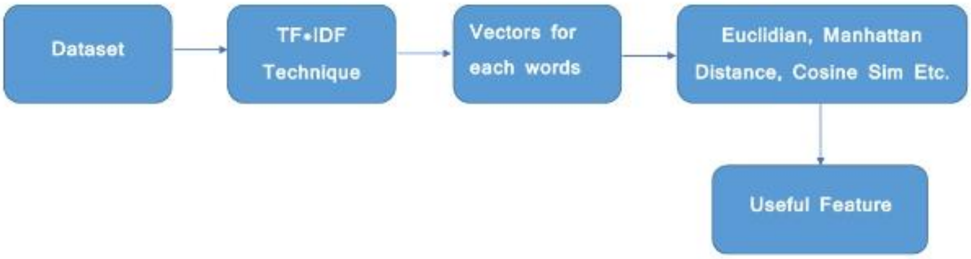


Fig. 2: TF-IDF block diagram.

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.

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

Step 4: Implementing TF-IDF: To make TF-IDF from scratch in python, let’s imagine those two sentences from different document:

first sentence: “Data Science is the sexiest job of the 21st century”.

second sentence: “machine learning is the key for data science”.

Natural Language Toolkit (NLTK)

NLTK is a toolkit build for working with NLP in Python. It provides us various text processing libraries with a lot of test datasets. A variety of tasks can be performed using NLTK such as tokenization, lower case conversion, Stop Words removal, stemming, and lemmatization.

Tokenization

The breaking down of text into smaller units is called tokens. tokens are a small part of that text. If we have a sentence, the idea is to separate each word and build a vocabulary such that we can represent all words uniquely in a list. Numbers, words, etc. all fall under tokens.

Lower case conversion

We want our model to not get confused by seeing the same word with different cases like one starting with capital and one without and interpret both differently. So we convert all words into the lower case to avoid redundancy in the token list.

Stop Words removal

When we use the features from a text to model, we will encounter a lot of noise. These are the stop words like the, he, her, etc… which don't help us and just be removed before processing for cleaner processing inside the model. With NLTK we can see all the stop words available in the English language.

Stemming

In our text we may find many words like playing, played, playfully, etc… which have a root word, play all of these convey the same meaning. So we can just extract the root word and remove the rest. Here the root word formed is called 'stem' and it is not necessarily that stem needs to exist and have a meaning. Just by committing the suffix and prefix, we generate the stems.

Lemmatization

We want to extract the base form of the word here. The word extracted here is called Lemma and it is available in the dictionary. We have the WordNet corpus and the lemma generated will be available in this corpus. NLTK provides us with the WordNet Lemmatizer that makes use of the WordNet Database to lookup lemmas of words.

Textblob

TextBlob is an open-source Python library that is very easy to use for processing text data. It offers many built-in methods for common natural language processing tasks. Some of the tasks where I prefer to use it over other Python libraries are spelling correction, part of speech tagging, and text classification. But it can be used for various NLP tasks like:

- Noun phrase extraction
- Part of speech tagging
- Sentiment Analysis
- Text Classification
- Tokenization
- Word and phrase frequencies
- Parsing
- n-grams
- Word inflexion
- Spelling Correction

Decision Tree Classification Algorithm

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question and based on the answer (Yes/No), it further split the tree into subtrees.

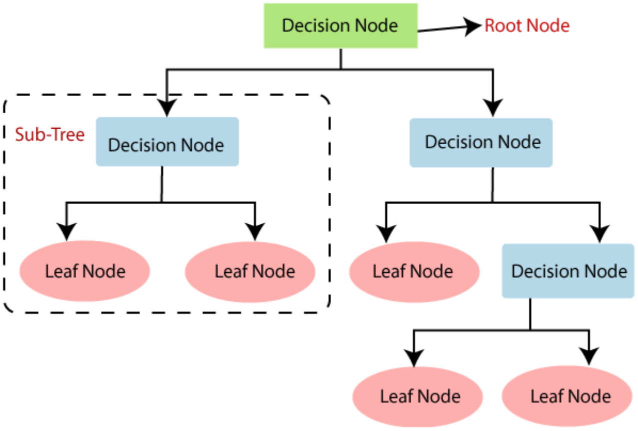


Fig. 3: General structure of a Decision tree.

Why use Decision Trees?

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while deciding, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies

- **Root Node**: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

How does the Decision Tree algorithm Work?

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step-3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example: Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer).

Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM. By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- **Information Gain**
- **Gini Index**

1. Information Gain:

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy (each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

- S= Total number of samples

- P(yes)= probability of yes
- P(no)= probability of no

2. Gini Index

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

$$\text{Gini Index}= 1- \sum j P j 2$$

Pruning: Getting an Optimal Decision tree

Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

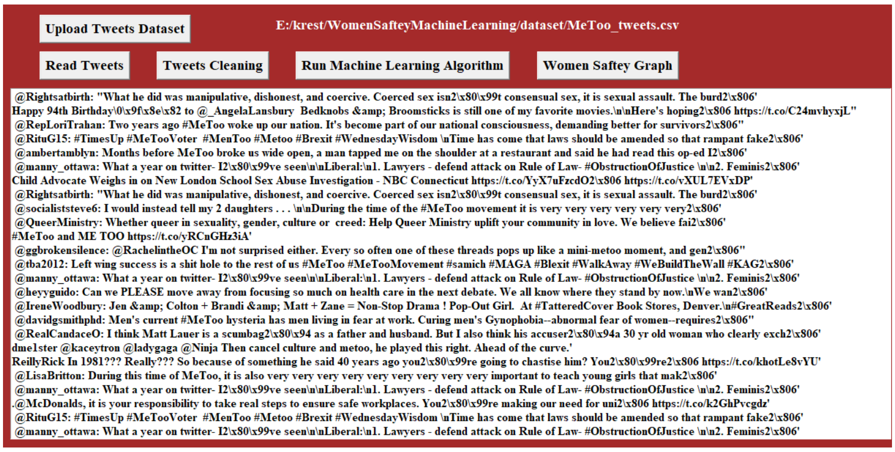
A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of tree pruning technology used:

- Cost Complexity Pruning
- Reduced Error Pruning.

Advantages of the proposed system

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

4. RESULTS AND DISCUSSION

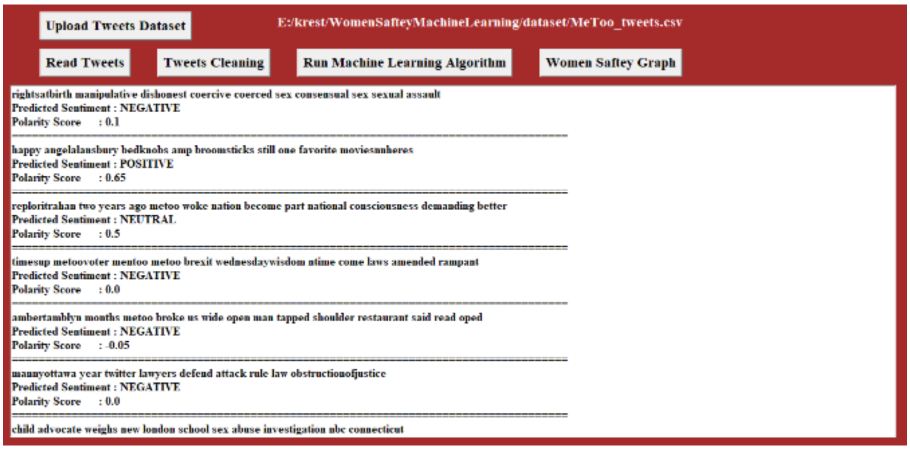


E:/krest/WomenSafteyMachineLearning/dataset/MeToo_tweets.csv

```
 @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'
```

Fig. 4: Sample dataset.

In Fig. 4 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



E:/krest/WomenSafteyMachineLearning/dataset/MeToo_tweets.csv

```
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
```

Fig. 5: Tweet sentiments with polarity score.

In Fig. 5 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 Safety 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.

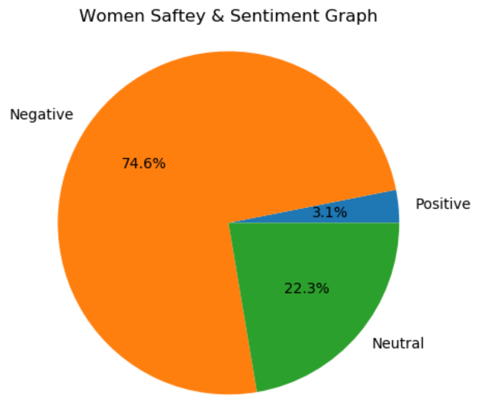
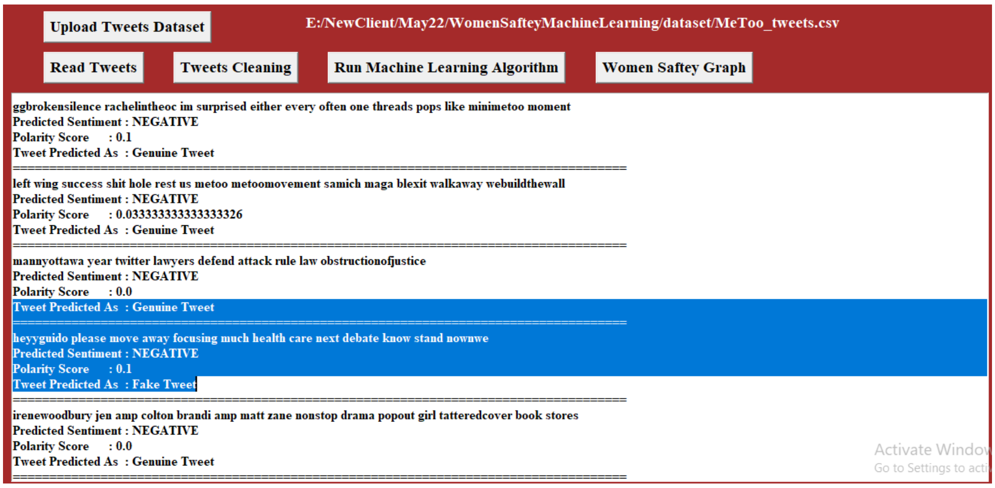


Fig. 6: Graphical representation of Women safety sentiment.

In Fig. 6 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 and neutral



E:/NewClient/May22/WomenSafteyMachineLearning/dataset/MeToo_tweets.csv

```
ggbrokensilence rachelintheoc im surprised either every often one threads pops like minimetoo moment
Predicted Sentiment : NEGATIVE
Polarity Score     : 0.1
Tweet Predicted As  : Genuine Tweet
================================================================================
left wing success shit hole rest us metoo metoomovement samich maga blexit walkaway webuildthewall
Predicted Sentiment : NEGATIVE
Polarity Score     : 0.033333333333333326
Tweet Predicted As  : Genuine Tweet
================================================================================
mannyottawa year twitter lawyers defend attack rule law obstructionofjustice
Predicted Sentiment : NEGATIVE
Polarity Score     : 0.0
Tweet Predicted As  : Genuine Tweet
================================================================================
heyyguido please move away focusing much health care next debate know stand nownwe
Predicted Sentiment : NEGATIVE
Polarity Score     : 0.1
Tweet Predicted As  : Fake Tweet
================================================================================
irenewoodbury jen amp colton brandi amp matt zane nonstop drama popout girl tatteredcover book stores
Predicted Sentiment : NEGATIVE
Polarity Score     : 0.0
Tweet Predicted As  : Genuine Tweet
================================================================================
```

Fig. 7: Predicted class of tweet.

In Fig. 7 we are displaying tweets with sentiment and its authenticity as FAKE or GENUINE by using decision tree algorithm

5. 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.

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] A. 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.