Crowdmapping on the Mitigation of Sexual Harassment

by

Rezwana Hasan
18101247
Ramisa Farhat Noor
18101557
Nabiha Noor
18101675

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Rezwana Hasan

Rezwana Hasan
18101247

Ramisa Farhat Noor
18101557

NabihaNoor

Nabiha Noor
18101675

Approval

The thesis/project titled “Crowdmapping o the Mitigation of Sexual Harassment” submitted by

1. Rezwana Hasan (18101247)

2. Ramisa Farhat Noor (18101557)

3. Nabiha Noor (18101675)

Of Fall, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 20, 2022.

Examining Committee:

Supervisor:
(Member)

Jannatun Noor Mukta
Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Faisal Bin Ashraf
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Sexual harassment is a gruesome act of violence that almost 84% of women in Bangladesh, according to Sajida Foundation [21], from the age of three to sixty, have endured it. With the use of crowdsourcing technology, our proposed idea has centralized towards the development of a system to bring down the statistics of such crime. Exploiting an open source crowdmapping platform along with Frontline SMS, an open source software used to collect information via text, to create a cartographic representation of incidents of sexual assault that are reported through multiple channels and data streams, including SMS texting, email and Twitter, while respecting the integrity of the victim. The objective of our vision is to be able to use this platform in the real world to overcome the cultural and environmental constraints that have hindered traditionally in collection of data for protection of women.

Keywords: Crowdsourcing technology, open source software, data streams.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor Jannatun Noor Mukta Miss for her kind support and advice in our work. She helped us whenever we needed help.

Thirdly, to our co-advisor Faisal Bin Ashraf Sir for his suggestions and immense support in our work.

Fourthly, the whole judging panel of 8th International Conference on Networking, Systems and Security (8th NSysS 2021). Our poster presentation was appreciated by them and all the reviews they gave us opened up new ideas for our further work.

And finally to our parents without their throughout support it would not have been possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

NLP Natural Language Processing

PCA Principal Component Analysis

PTSD Post-traumatic stress disorder

WCSS Within-Cluster Sum of Square

Chapter 1

Introduction

1.1 Introduction

According to [6] “Sexual harassment is intimidation, bullying or coercion of a sexual nature, or the unwelcome or inappropriate promise of rewards in exchange for sexual favors.” Sexual harassment is an objectionable act that one does to another without his/her permission. This distasteful act includes, according to [30], a stranger or even someone known to the victim, touching, groping or making any sort of physical contact to the victim, without the victim’s consent. It also includes catcalling, eve-teasing and one creating a relationship with another only for sexual favours, keeping the victim in the unknown. Someone showing interest in another person’s sex life and making that person feel uncomfortable is also a part of sexual harassment. It also includes making sexual comments on someone and making sexual gestures towards someone. Cases of sexual harassment are also becoming increasingly visible in social media. Teasing a person, blackmailing them with personal facts or photographs, making sexual comments and other such cyber-bullying offences are rising day by day.

The most dreadful form of sexual harassment is when an individual is raped. According to [25] , rape is a sexual activity which is a criminal offence, that includes unwilling sexual intercourse with the victim. The victim, in this case, is compelled to undergo this unpleasant act. Usually this crime occurs to individuals who are frightened to raise their voices, are a minority in the society, are suffering from some sort of mental disability, which is why they are unable to give consent and at the same time are unaware of the severity of the deed. Some victims are also under the influence of drugs or unconscious while this loathsome act is performed on them. The cause of rape used to be thought of as a result of uncontrolled sexual desire, but nowadays, it is considered to be the cause of persistent power imposition on the victim.

1.2 Problem Statement

Sexual harassment is a common affair in the entire world. Sexual harassment is not a new occurrence. It has existed in society for a very long time. However, according to [2] , Bangladesh has very recently started to acknowledge the need to take action against such loathsome happenings. Even though many laws already existed, which were specifically against sexual harassment, Bangladesh has very recently started to become intolerant against this act, being impacted by the MeToo Movement which spread across the world in 2017 [14].

Whereas a good number of the population of Bangladesh do understand the severity of the act and are protesting it, there are still quite a large number of the population who still have an incorrect view on this topic because they are unaware of the laws regarding sexual harassment [21]. The article also quotes, "About 84% of women in Bangladesh are constantly being sexually harassed on the road, in vehicles, in educational institutions, at work, and even at home." The sexual harassment cases against women have been rapidly increasing in Bangladesh. Women are mostly being harassed in public places, buses, shopping centres, streets and many other places. Even during the Covid-19 pandemic, as mentioned in [24] . In the case of women, it has been seen that women feel afraid to raise their voices, as a victim, mainly in fear of becoming a matter of humiliation and shame for the society. Also, the victim is unaware of the rules and regulations of reporting a case. Some also fear that they would lose their jobs which would greatly affect their finances. If a report is made, the victims are also scared of the fact that the matter would not be handled privately and the entire society would come to know of it. According to [26], there are also such cases where the women would end up holding themselves responsible for their dreadful experience and some would not do anything about the situation believing that no action will be taken. Sexual assaults on women are not only committed by strangers, but also by their family members or friends, which is why some women end up not reporting about the assault.

According to [18], a research done by the United Nations stated that approximately 3.7% of men in rural Bangladesh raped another man. Around 20 boys experienced sexual harassment from January to September in 2020, this statement was released by the rights organization Ain o Salish Kendra. In that very article it also mentions that Prof Dr Tania Haque stated that the rapes or sexual assaults of boys or men are not disclosed due to the "side effect of masculinity". Due to this harmful and destructive view of men and their masculinity, the men think that if they admit that they were subjects to sexual violence they may face public humiliation and identity crisis. Therefore, a lot of men or boys do not come forward and stand up for themselves. Due to this reason, no reliable research has been done yet. Moreover, no one has come up with an effective solution to combat this issue.There are innumerable challenges when it comes to solving a sexual harassment case in Bangladesh.

Sexual harassment has a huge impact on the victim. The trauma that the victim goes through is unimaginable. [9] states that the victims seem to separate themselves from the rest of the world and create a small world of darkness for themselves which consists of anxiety, trauma, depression and stress. The victim loses self confidence. The victims might go through a lot of trauma which might affect them for a short period of time or may last for a long time, depending on how strong the victim is emotionally. The victim may also undergo a lot of physical pain like bleeding, difficulty walking, broken bones, unwanted pregnancy, sexually transmitted diseases and many more. The trauma of the event might also lead the victim to think of suicidal attempts. Post-traumatic stress disorder (PTSD) like having recollection of the act is likely to continue for a long period of time.

According to [14], “Section 509 of the Penal Code 1860 criminalises acts, words and gestures intended to “outrage the modesty of a woman” with a prison sentence that may extend to one year along with fines.” In this statement, as mentioned by daily star, there is an unnecessary mention of woman’s modesty, where as this should not even be a factor in the scenario of sexual harassment. Thus, the law itself is victimising the women rather than the assaulter, which in turn takes away the feeling of being safeguarded from the women. An offence has been introduced by the Nari-O-Shishu Nirjatan Ain 2000 (section 10) which is known as jounopiron which is also known as sexual oppression. This can help illegalize and criminalize the act of touching any woman or child with any part of the criminal’s body or with an object in order to illegally satisfy their sexual desires.

1.3 Aims and Objectives

While the media only cover the extreme forms of sexual violence like rape and domestic violence, the research on related papers [3], [8], [28] show that persistant and pervasive acts of street harassment can affect women’s security, freedom and safety in public places.

The main objective of this research paper is to build a source to mitigate the cases of sexual harassment in public places. This paper demonstrates a platform to be used as an open source to collect data of harassment experienced by women in public territories respecting the integrity and identity of the victims. This paper also describes the implementation of such a platform to be used with ease in daily life.

1.4 Research Methodology

We are building a web application that keeps a record of sexual harassment cases that have occurred in Bangladesh. In this webapp, there is a map which portrays the statistics of such cases location wise. The locations are marked as red zones. The cases are mentioned in the map through the inputs that the users are providing to the webapp.

The users can insert the input through crowdsourcing technology via an open source software. The user has the authority to mark the zone while keeping the identity and integrity hidden. The software now keeps the records of the cases.

In order to carry out the entire process, a Machine Learning model has been used. Below, in Figure 1, the flowchart representation of the entire model that we used is shown.

At first, an input dataset is provided to the model. The type of learning that is being used in this process is unsupervised machine learning. In this type of learning, unlabelled data is provided to the model. Machine learning algorithms are used to cluster and analyse the dataset. According to [18], the algorithm applied to the model then provides a labelled output. The labelling is done based on the similarities and the differences in the data.

So, as the input training data is fed into the model, it undergoes preprocessing. It prepares the input data in such a way that the machine learning algorithm can easily process it. The preprocessed data undergoes processing. The data is clustered into groups based on their similarities or differences, which are given labels after clustering is complete.

Next, we compare the labelled outputs produced after clustering with the expected output. If the labelled output matches the expected one, then the model is a success. Otherwise, it is a failure.

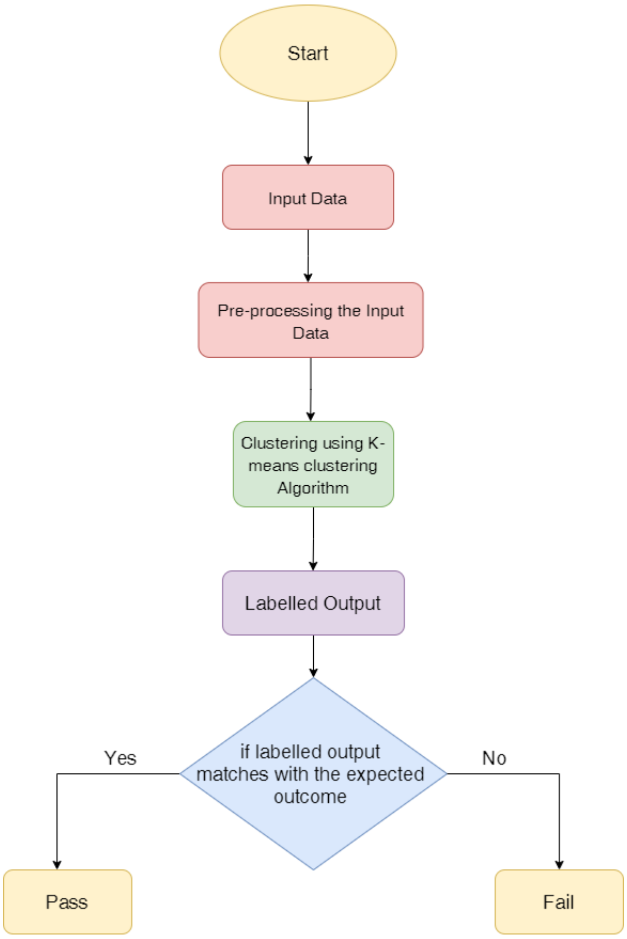

Start
Input Data
Pre-processing the Input Data
Clustering using K-means clustering Algorithm
Labelled Output
if labelled output matches with the expected outcome
Yes
No
Pass
Fail

Figure 1.1: Flowchart of the Machine Learning Model

1.5 Thesis Outline

The thesis work is concerned with detecting and reducing sexual harassment cases that occur in the streets.

The first part (Chapter 1) of the paper is the introduction where the rising occurrences of sexual harassment cases have been mentioned and how it is not given that much importance as it should be given. The law of Bangladesh is not strong enough when it comes to a victim who has been sexually harassed. Finally, it states our contribution in reducing the cases of sexual harassment and the work flow that would be followed.

Secondly, in the part which discusses related work (Chapter 2), a few related papers have been mentioned which carry the same objective as ours.

The third chapter (Chapter 3) of this paper discusses the data collection procedure and how a machine learning model was selected to carry out the work. It gives brief summary of the way data was collected from the respondents and also mentions how the validation of the dataset is not possible to be met. It also discusses why the model selected was unsupervised machine learning model.

The fourth chapter (Chapter 4) mentions how the data from the dataset was preprocessed. The raw data was modified into the sort which would be easily fed into a machine learning model. The reason why label encoding was carried out instead of one hot encoding is mentioned.

Chapter five (Chapter 5) is the most important chapter as it contains the data analysis and results. It describes how the data were processed, firstly using the elbow method to find the number of clusters and then K-means clustering is used to form clusters from the dataset. It discusses why K-means algorithm was used. Decision tree classifier is used to identify the pattern of the clusters. It also shares why Hierarchical clustering algorithm is not preferred. Later it talks about how the text data were processed using vectorization and PCA.

Finally, the last chapter (Chapter 6) discusses the entire thesis work that has been done and the findings that have been made. It concludes the thesis work and shares the future work that can be carried on.

Chapter 2

Related Work

As the public vulnerability towards women is no longer a new thing, steps must be taken to diminish such gruesome acts. There has been much research over the past few years to support the integrity of women's safety outside their homes.

A research paper [9] portrays an idea of a platform on the crowdsourcing based tool called "HarassMap" in Egypt. This uses an open source software developed by Ushahidi for mapping and reporting the incidents around. This platform helps to overcome the hesitation that women usually feel to report their experienced incident due to the shame and social constraints. This paper also ensures the security and integrity of the victim.

Another similar paper [10] demonstrates the implementation of Ushahidi on the HarassMap interfacing with the Google Map. This application allows the platform to have an aerial target of the landscapes and then build a visual cartography.

This platform is quite popular and an adaptable method for women in Egypt and other Arab countries where the demand for safety of women is often overlooked. Women can report any type of harassment, keeping their identity hidden, through this platform which Ushahidi then collects the data to connect in real-time to the interactive map. These complaints are then geo visualised as geographical data points on the Google Maps in the form of a red dot that can be added based on the position of the zoom toggle. This produces a map marked with red dots to indicate areas which are prone to violence and harassment. This map can be viewed by anyone to alert them about the red zone areas.

Even though this project has gained praise over the years, there are still setbacks surrounding the implication. [9] suggests that few reports collected could be inaccurate. False data or reports could be added into the platform by any individual or a group of people to target a specific neighbourhood.

[10] also defines that aerial observation made from different points of view can result in distorted images as the identities could be far from the objective.

[8] tries to solve this situation by developing a web and mobile phone based application called "Protibadi". This is designed to report and share women's stories around

sexual harassment in public places.

In this project, there is a tab called “Save Me” which is used as a panic button. Whenever any person is feeling uncomfortable, that person can press that button. After pressing this button, a loud noise from the mobile phone’s speaker is emitted which draws public attention in that zone. Simultaneously, a text message containing the location of the user(with the help of GPS) will be sent out to each of the emergency contacts which will also let the contacts know that the user is in trouble.

Another tab is included, so that the user can edit, add or delete the emergency contact list. Lastly, there is another tab which is included for reporting. The user can report any incident of harassment that he/she experienced or witnessed.

Even though this system is a great initiative towards the problem every other woman is dealing with every single day, there are some situations or environments that the designers did not take under consideration while developing the system. One of the circumstances overlooked by this paper was that when the user is in danger, it can be difficult for her/him to reach out to their phone and find the right tab. Moreover, if the device is stolen before the user could even find the right tab then the application will be of no help.

This study is not only inspired by the approaches extended by the paper mentioned above but also tries to combat the overlooked unfortunate events that could happen in any situation. Features like- the user being able to choose the route after seeing the amount of “red zone” in a path, the automated share location feature after the victim enters the “red zone”, the option of submitting any kind of evidence like -video recording, picture or voice recording while reporting any sort of harassment experienced or witnessed by the user make our system more efficient, distinct and practical.

Just like this paper [12] suggests, there was another platform introduced known as Safecity.in which was used in India. Safecity.in was introduced in 2012 in protest of a case where a woman was gang-raped and her boyfriend was beaten on a bus, on their way home, returning from the movie theatre. The woman took her last breath right after that brutal incident. [7] Safecity.in is a platform that allows to reduce violence in public and private areas. It uses crowdsourcing to gather information from the victims, whose identities are kept private. This platform helps to make the citizens aware of any public violent incidents, upgrade policies with the help of data that are being collected and help to effectively set the budget for proper resource allocation.

Safecity.in, likewise, takes inputs of the cases from the victims and portrays the “hot spots” of those incidents in a map. It records the date and place of the incident. The validity of the incidents is also monitored by keeping track of the areas where the incidents occur and the types of incidents. These later help to trace patterns,using grounded theory, and confirm how valid and reliable the data collected is. The platform also keeps record of the victims’ responses during such occurrences: if the victim did not raise her voice, if she escaped or if she fought back.

This research paper mainly focuses on sexual harassment cases, without considering the gender, whereas that paper using Safecity.in solely focuses on sexual harassment cases on women. A limitation noticed in the paper was that the questions were open-ended, i.e. the victims are allowed to write what they want without any boundaries. The question pattern is not as such where the victim has to select from a set of possible answers. This narrows down the scope of collecting each and every required detail about the incident as the victim might miss few of the details.

Chapter 3

Data Collection and Model Selection

This thesis work aims to build a web application that keeps a record of sexual harassment cases that have occurred in Bangladesh. The web application will consist of a map which will portray the statistics of such cases location wise. Through the users' inputs, provided through the web application, the locations where the harassment cases occurred are marked. The users can insert the inputs through a crowdsourcing technology via an open source software. The users can keep their identity hidden during this process. The software, hence, keeps the records of the cases.

3.1 Data Collection

Sexual harassment is a very common issue nowadays. It is of no surprise that almost 84% of the women in Bangladesh have faced sexual harassment,according to Sajida Foundation. Not only women, but also men, children and transgenders face sexual harassment in their daily lives. In order to collect information about such sexual harassment cases, an online survey was conducted. The choice of collecting data is via online surveys because most people now have access to the internet which makes it an easy platform to collect and assemble the raw data successfully from a large group of people. The survey form, consisting of questions built on the Google Form, was sent out to people via social media. Hence the dataset that will be acquired would be primary data.

The survey was done among people of all genders and age. The details of the sexual harassment cases like: the incident, location, time and date of the incident and the transportation used by the victim were collected. It was also collected if the victims were able to come out of the unpleasant scenario safely or not. The respondents could respond to the survey as both a victim or as a witness.

The data collection was done anonymously. Keeping the identity of the respondents hidden was obligatory. When it comes to a sensitive and intimate topic such as sexual harassment, individuals prefer not to reveal their identity. The victims usually fear public exposure, social shame and abuse. The victims or their families may also get threats from the assaulter, which leads to a series of miserable events. The witnesses are also sometimes not confident enough to speak up about the incidents, that they have witnessed, due to the fear of being targeted. Hence, the emails or any personal information of the respondents were not collected.

The validation of the data collected cannot be verified due to the anonymity of the data collection process. According to [5] ” The provision of complete anonymity is presumed to facilitate collection of more accurate data by minimizing social desirability pressures.”

Collecting data was the initial task for this research to coax into the pattern of prevalence of such incidents across the country. As soon as the people took part in the survey, the number of the responses and their responses were recorded on Google Excel sheet.

3.2 Unlabelled Data

The responses recorded on Google Excel Sheet include the age and gender of the victim, the date, time and location of the incident that occurred and which transportation was used by the victim. From these collected data, the aim is to find out the locations where sexual harassment cases are occurring in large, medium and small numbers. Hence, the dataset generated from the collected data is unlabelled data. Unlabelled data are the pieces of information which are not put under any characterization or classification.

3.3 Unsupervised Machine Learning

Since the responses are being recorded manually by the people, this raw dataset requires processing to predict a pattern of the weight of sexual harassment across the country. According to [4] The dataset is unlabelled, which is why unsupervised machine learning is used. As mentioned [20] Unsupervised machine learning finds all sorts of unspecified patterns in the data and helps to discover features which come to use during classification. Unsupervised machine learning utilises machine learning algorithms to find the patterns in the data, which allows them to form clusters, without any human involvement. Thus, unsupervised machine learning makes it easier to explore the dataset, resulting in discovery of similarities and differences in the information to form clusters.

Chapter 4

Data Preprocessing

The data collected through online survey is the raw data that needs to be processed before feeding it into the machine learning model. It is termed as raw data because this data needs to be prepared and structured into a set of clean data which can be used for further processing. According to [1] Data preprocessing is the process by which these raw data can be transformed into clean data, which can then be provided to the machine learning model.

While creating the questionnaire for the online survey, the location of the incident had to be divided into 3 parts: Road, Area and District. This was done for the ease of data collection, otherwise each respondent would have filled the location in a different way, which would make it difficult to process. Road, Area and District are recorded in three columns, which are then merged in to one column, known as location. Then those three columns were dropped.

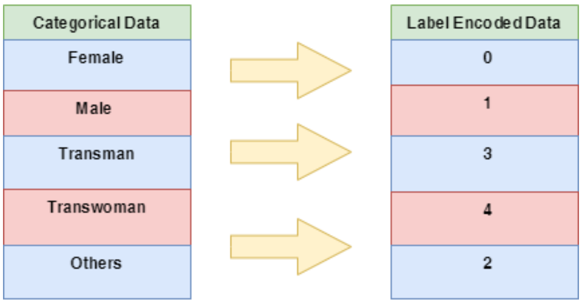
The columns which included text data: ”Detailed Description of the Incident”, ”How did the victim handle the situation? If the victim did come out of the situation safely, please mention how.”, ”Did you report the incident? If yes, how did they respond?” were dropped for extraction of categorical data, so that the rest of the data could be processed and labels could be generated from the dataset.

4.1 Label Encoding

Label encoding is a popular encoding technique for dealing with categorical data. Based on alphabetical classification, each label is issued a unique integer in this manner. Categorical data is data that expresses “qualitativeness” through a string of words. As the variables used in this study such as ”Gender”, ”Transportation utilized by the victim”, categories of event, description details of the occurrence, and so on are descriptive and qualitative, the categorical dataset used in this study is heavily dominated by Nominal Data. Because the computer only understands numbers, label encoding labels this data with dummy numbers, allowing it to be preprocessed for future processing. Here’s an example of how data is encoded.

One hot encoding technique is another encoding method. One hot encoding is a method of encoding in which each categorical feature is represented by binary values of 1 or 0. One hot encoding is preferred over label encoding because a priority is assigned by default in the label encoding procedure. When the number of categorical features is low, one-hot encoding is favoured since it will be more effective. However, the dataset utilized in this work has a large number of categorical variables, making it challenging to analyze and extract rules for detecting the label that this unlabeled data requires. Furthermore, the categorical data in the utilized dataset is not ordinal, which is the ideal data for label encoding.

This is why label encoding is used with the dataset since it will make processing easier.



Categorical Data	Label Encoded Data
Female	0
Male	1
Transman	3
Transwoman	4
Others	2

Figure 4.1: Label Encoding

Chapter 5

Data Analysis and Results

5.1 Data Processing

5.1.1 Elbow Method

The Elbow Method is a process which is typically used to detect the best fitted number of clusters the given dataset to divide into.The value of ”K” the optimal number of clusters recommended by the method. The value of K ranging from 1 to 10 is taken. [13] for every value of K, the sum of squared distance between each point and the centroid in a cluster is being calculated(calculates the sum of the square of the points and calculates the average distance.) which is known as WCSS (Within-Cluster Sum of Square) . After plotting the sum against the value of K, it can be seen that with the increment of the number of clusters, the sum decreases.At one point the plotting shows an “elbow” like point on that graph, the value of K at that point is the optimal value fitted for this model.

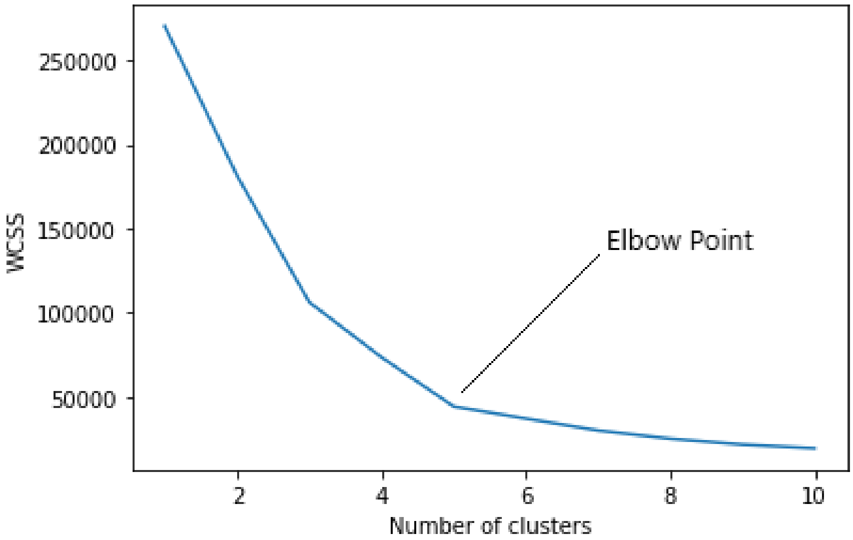


Figure 5.1: Elbow Point Example

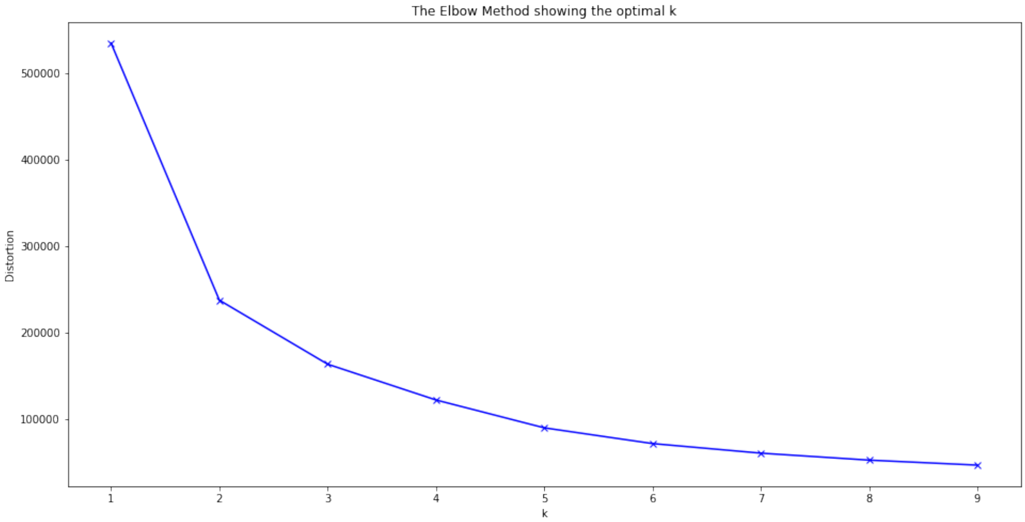


Figure 5.2: Elbow Method giving optimal value of k

5.1.2 K-means Clustering

[11] explains clustering is a technique that involves grouping a set of objects in a way that they are similar to each other. It helps to organize the data from the dataset in such a way which portrays the internal structure of the dataset. One of the famous techniques for clustering is the K-means clustering algorithm. As the dataset used in this paper is unlabelled, the K-means algorithm is more suitable.

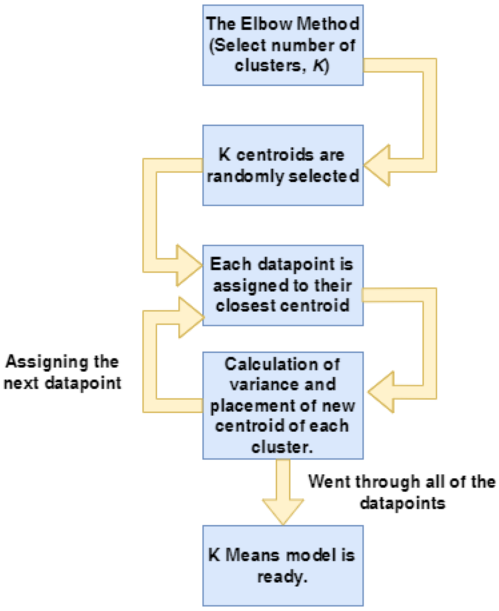
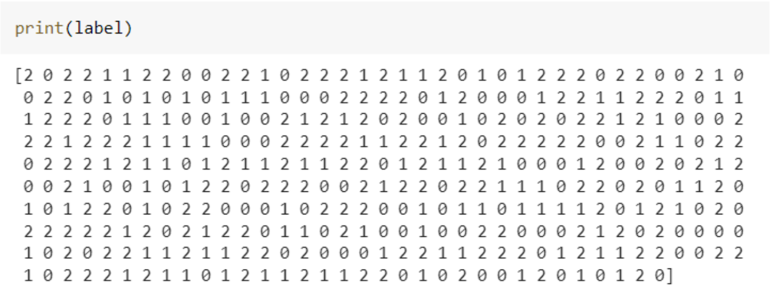

The Elbow Method (Select number of clusters, K)
K centroids are randomly selected
Each datapoint is assigned to their closest centroid
Assigning the next datapoint
Calculation of variance and placement of new centroid of each cluster.
Went through all of the datapoints
K Means model is ready.

Figure 5.3: How K-means clustering works.

After getting the value of K from The Elbow Method, now the K-means clustering method can be implemented. With this value of K , a more precise grouping in this dataset is possible through the K-means clustering. K means clustering separates the unlabeled dataset into k groups, with the dataset in each group having comparable characteristics. Every cluster has a centroid, which is also known as the cluster's multidimensional mean. As a centroid-based method, this procedure aims at reducing the distance between each data point and its own cluster.

The form of the retrieved dataset used for processing is (366,5), which is reduced to two dimensions using PCA, or the Principal Component Analysis algorithm, for a better understanding of the set of data. It is intended to improve readability while minimizing data redundancy.

Then K means was applied on the dataset used in this paper, where all of the data points were divided into groups with other data points with similar properties.



```
print(label)
```

```
[2 0 2 2 1 1 2 2 0 0 2 2 1 0 2 2 2 1 2 1 1 2 0 1 0 1 2 2 2 0 2 2 0 0 2 1 0
 0 2 2 0 1 0 1 0 1 0 1 1 1 0 0 0 2 2 2 2 0 1 2 0 0 0 1 2 2 1 1 2 2 2 0 1 1
 1 2 2 2 0 1 1 1 0 0 1 0 0 2 1 2 1 2 0 2 0 0 1 0 2 0 2 0 2 2 1 2 1 0 0 0 2
 2 2 1 2 2 2 1 1 1 1 0 0 0 2 2 2 2 1 1 2 2 1 2 0 2 2 2 2 2 0 0 2 1 1 0 2 2
 0 2 2 2 1 2 1 1 0 1 2 1 1 2 1 1 2 2 0 1 2 1 1 2 1 0 0 0 1 2 0 0 2 0 2 1 2
 0 0 2 1 0 0 1 0 1 2 2 0 2 2 2 0 0 2 1 2 2 0 2 2 1 1 1 0 2 2 0 2 0 1 1 2 0
 1 0 1 2 2 0 1 0 2 2 0 0 0 1 0 2 2 2 0 0 1 0 1 1 0 1 1 1 1 2 0 1 2 1 0 2 0
 2 2 2 2 2 1 2 0 2 1 2 2 0 1 1 0 2 1 0 0 1 0 0 2 2 0 0 0 2 1 2 0 2 0 0 0 0
 1 0 2 0 2 2 1 1 2 1 1 2 2 0 2 0 0 0 1 2 2 1 1 2 2 2 0 1 2 1 1 2 2 0 0 2 2
 1 0 2 2 2 1 2 1 1 0 1 2 1 1 2 1 1 2 2 0 1 0 2 0 0 1 2 0 1 0 1 2 0]
```

Figure 5.4: Generated class names

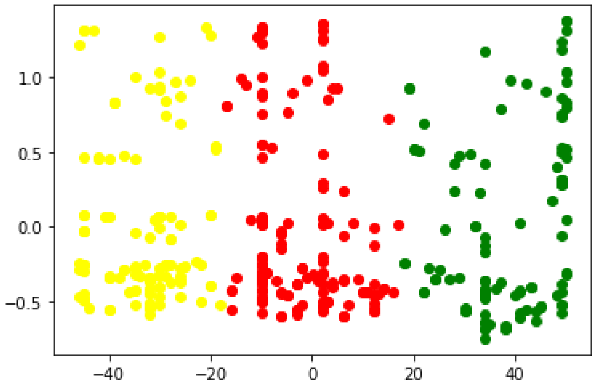


Figure 5.5: K-means clusters

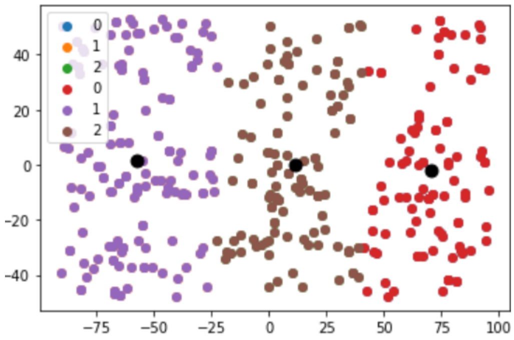
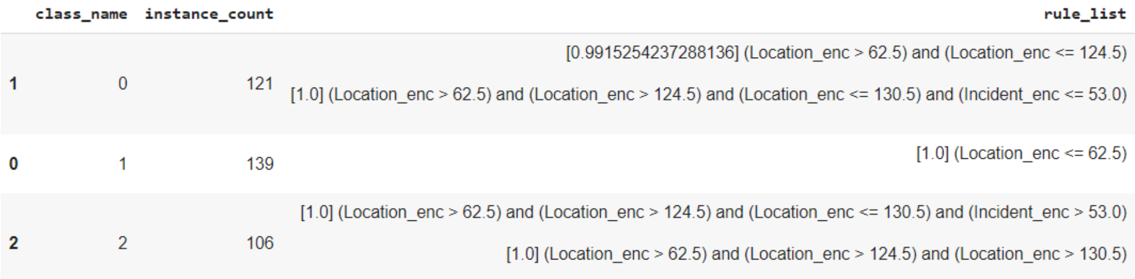


Figure 5.6: K-means clusters with centroids

Why K-means Clustering Algorithm?
K-means clustering algorithm is preferred to find clusters for this dataset because it is an uncomplicated algorithm in terms of execution. [27] The dataset being used is a vast dataset and K-means algorithm can handle vast datasets and can easily adjust with different new entries. Since the variables of the dataset are large, choosing a smaller value of "K" helps in reducing time complexity. Also, [15] K-means clustering algorithm forms hard clusters. Hard clustering combines data points in such a way that a single cluster contains each data point. So, the K-means clustering algorithm is a centroid-based algorithm where each data point is allocated to its nearest centroid, thus forming tight clusters.

5.1.3 Decision Tree Classifier

The next step would be to identify the pattern of the cluster and how the zones are grouped in each cluster. To achieve that, a decision tree classifier is being used. In machine learning, a decision tree is a predictive modeling strategy. A binary recursive partitioning procedure is used to create such a tree. This is a method of splitting the data into segments and then dividing it on another branching. Decision tree classifier is typically used on supervised learning, to predict the target variables of the dataset. In this study, it was used to find the pattern of clustering and how the grouping is created and interpret the type of zone that the data point belongs to by considering the class generated by the clustering method as the target variables.



	class_name	instance_count	rule_list
1	0	121	[0.9915254237288136] (Location_enc > 62.5) and (Location_enc <= 124.5) [1.0] (Location_enc > 62.5) and (Location_enc > 124.5) and (Location_enc <= 130.5) and (Incident_enc <= 53.0)
0	1	139	[1.0] (Location_enc <= 62.5)
2	2	106	[1.0] (Location_enc > 62.5) and (Location_enc > 124.5) and (Location_enc <= 130.5) and (Incident_enc > 53.0) [1.0] (Location_enc > 62.5) and (Location_enc > 124.5) and (Location_enc > 130.5)

Figure 5.7: Decision Tree Classifier

After the decision tree classifier was applied to the dataset, pattern recognition of clustering becomes easier. As it can be seen, the response of 366 was divided into 121,139 and 106 grouped with 3 classes which can ultimately be considered as 3 types of zones(High, Low, Medium). And the primary deciders of this division was the location variable and the incident variable. In the above analysis, it is shown that in class 1, 139 responses were grouped and 100% of the locations are valued less or equal to 62.5. In class 0, 121 responses were grouped where around 99% of the locations' encoded value is more than 62.5 and less or equals to 124.5 and 100% of the time, the locations' encoded value is more than 62.5 and 124.5 and less or equals to 130.5 and the encoded value of incident is less or equals to 53.0. Similarly, the rest of the pattern can be detected.

Now, the concern arises that among all the features and variables , which variables are the major contributors in the whole experiment. To identify that, Random Forest was applied.

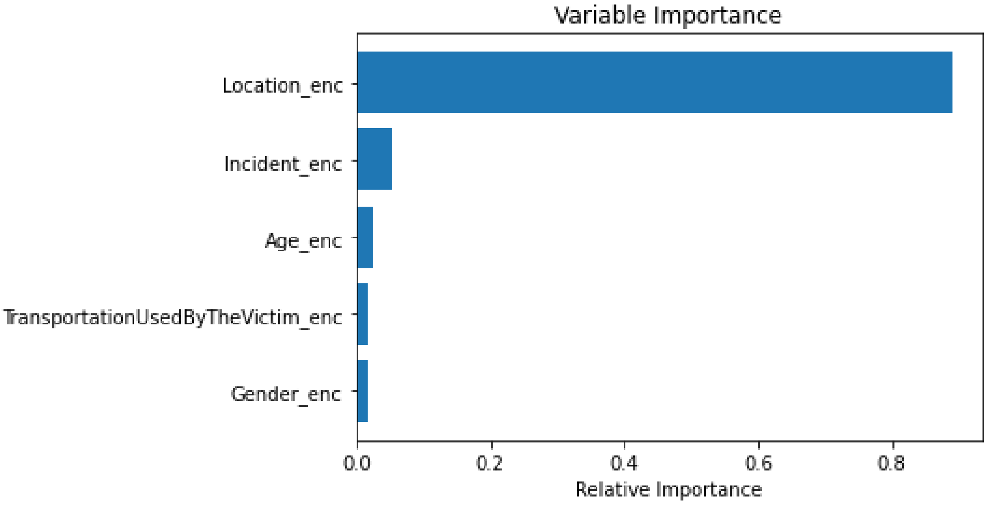


Figure 5.8: Feature Importance using Random Forest

After analyzing the importance report, the location variable is the most important factor in this dataset and contributes to the clustering method. And the second most important feature is the incident that happened during the harassment. Even though other features such as age, gender, time, transportation used by the victim were processed in this experiment, the most important factors were the location and incident of harassment.

5.1.4 Other Clustering Algorithms

Hierarchical Clustering Algorithm

Hierarchical clustering, also known as Hierarchical Clustering Analysis, can form clusters from an unlabelled dataset. In this clustering algorithm[23] each data point is taken into consideration as different clusters. The algorithm recognises the clusters which are adjacent to each other at the minimum distance (taking Euclidean distance into account) and unites two clusters which are mostly identical. This process of finding the closest and the most identical clusters proceeds until all the clusters are combined together. As a result a dendrogram is created which portrays the hierarchical relation between the clusters.

Even though the hierarchical clustering algorithm is suitable for unlabelled data, it is not preferred for this dataset. Hierarchical clustering algorithm is not suitable for a large dataset. [19] In this algorithm, there is no returning back when clusters have already been produced. If an erroneous cluster has been produced during the initial stages then the whole algorithm will carry out this error and hence the end result would be erroneous as well. Also there are different ways in which the similarity of two clusters can be determined. Each of these methods have their own drawbacks which sums up to a faulty outcome. Some of those drawbacks include: Complication in handling clusters of various sizes, being easily affected by outliers and noise and being problematic towards large clusters.

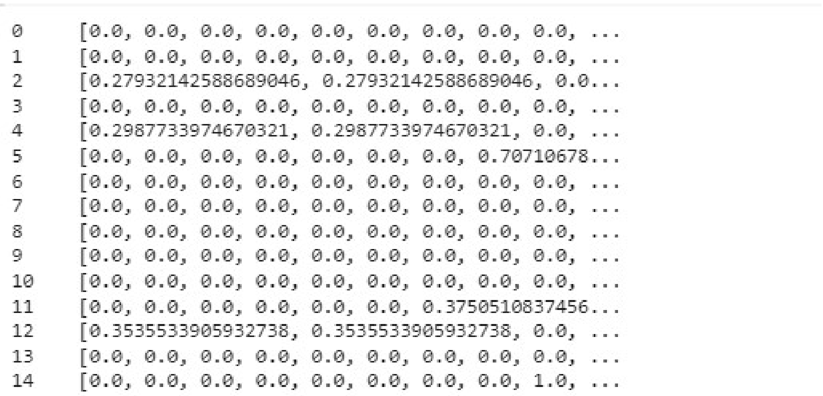
5.2 Text Preprocessing

As the dataset contains text data, there is a need to convert the texts in numerics for the machine. [22] So text preprocessing is the best method to clean the text from various noises. This eliminates the punctuations, emotions in human languages and is ready to be converted to numbers.

5.2.1 Vectorization

The cleaned words now are ready to be converted into vectors. One of the suitable methods for vectorization is Term Frequency-Inverse Document Frequency(TF-IDF) which is a statistical measure of relevance and frequency of the words in a document.[17] TF-IDF is calculated by multiplying two matrices, Term Frequency(TF) and Inverse Document Frequency(IDF).

After applying the vectorization on the cleaned text



```
0     [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
1     [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2     [0.27932142588689046, 0.27932142588689046, 0.0...
3     [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
4     [0.2987733974670321, 0.2987733974670321, 0.0, ...
5     [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.70710678...
6     [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
7     [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
8     [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
9     [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
10    [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
11    [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.3750510837456...
12    [0.3535533905932738, 0.3535533905932738, 0.0, ...
13    [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
14    [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, ...
```

Figure 5.9: Vectorization of Incident

This is one of the features where vectorization has been applied. The vectors consist of values 0s and other decimal figures. The values are in accordance with the importance of the words in the original text. The decimal figures represent the weight of the particular words in the text whereas the 0s show the rest of the words in the text. This gives a semantic meaning to the process.

5.2.2 PCA

Vectorization on different columns gives out different dimensions according to the semantic meaning of the text. Principal Component Analysis is applied on the vectors to reduce the dimensions into 2D planes. This technique is done to bring out variation and strong patterns in the dataset [29].

Detailed Description with Gender of the Victim

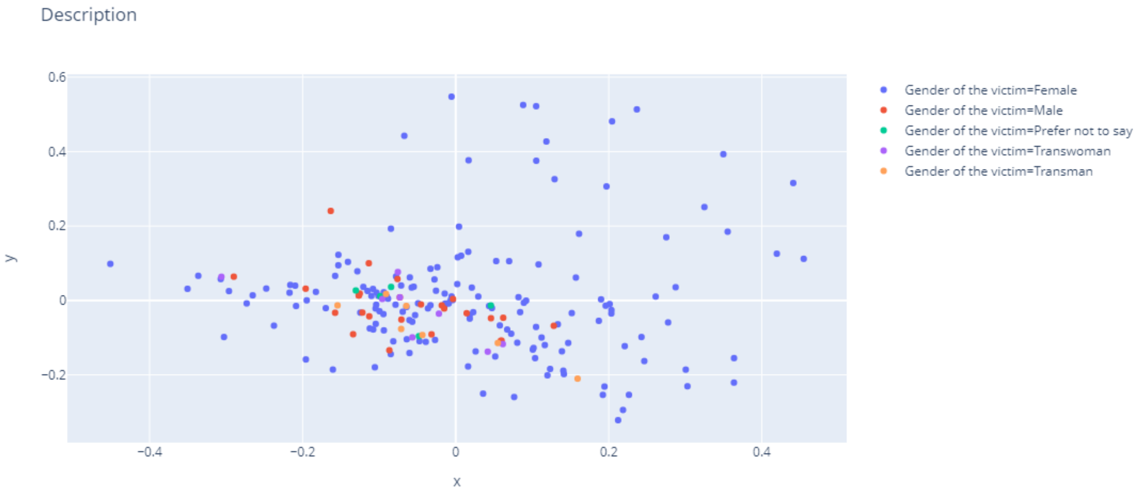


Figure 5.10: PCA on Detailed Description

In Figure 5.9, PCA is applied on the column 'Detailed Description' and color coded on the column 'Gender of the Victim'.

The highly correlated points are clustered together around the origin axis and on the first quadrant. This means on average, Female are the victims of most of the description and eventually they are positively correlated.

Incidents with Gender of the Victim

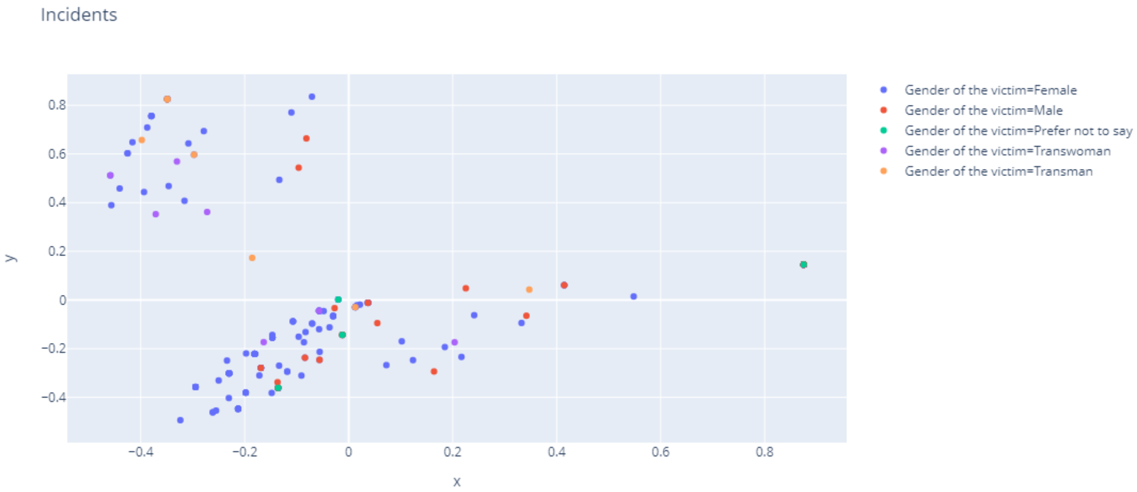


Figure 5.11: PCA on Incidents

Figure 5.10 shows PCA is applied on the column ‘Incidents’ and color coded on the column ‘Gender of the Victim’.

The highly correlated points are clustered together along with outliers. The outliers represent that those points are less in numbers with much variance in them.

Date and Time with Gender of the Victim

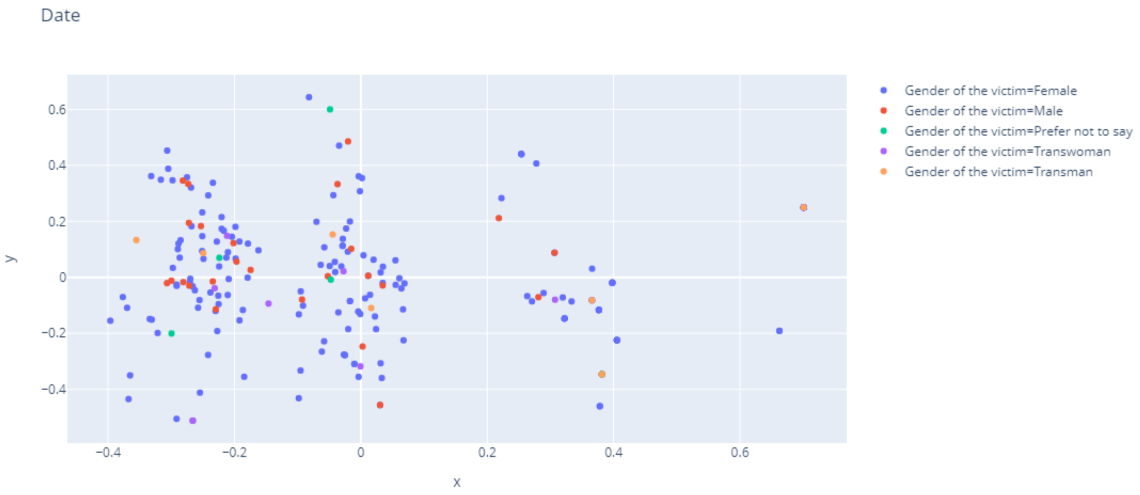


Figure 5.12: PCA on Date

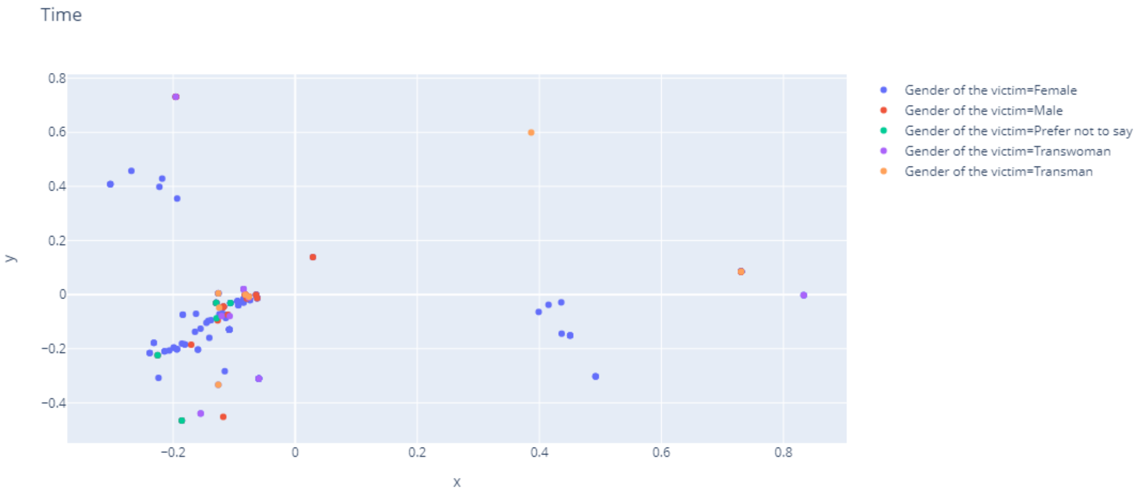


Figure 5.13: PCA on Time

In Figure 5.11, PCA is applied on the column ‘Date of the incident’ and color coded on the column ‘Gender of the Victim’.

In Figure 5.12, PCA is applied on the column ‘Time of the incident’ and color coded on the column ‘Gender of the Victim’.

In both of the above cases, the highly correlated points are clustered together which shows the weight of the information on the gender of the victim.The difference along x- axis represents more variance than the difference along y- axis. Since there is not much difference along the x-axis for the variable ‘Female’ in both of the graphs, ‘Female’ are the victims irrespective of time and date.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The data collected through various channels received around 360 answers. However, the validity of the acquired data would be questioned because of the need to protect the respondents' identity. After gathering the data, which is largely categorical in nature, it is preprocessed using label encoding. Label encoding was chosen since other encoding methods, such as one hot encoding, can greatly increase the dimension and quantity of features, making interpretation of that grouping very difficult. Label encoding was used to avoid this situation.Following preprocessing, the dataset is subjected to the Elbow technique, which provides the ideal value of K, the number of clusters or groups into which the dataset should be divided. For the given dataset, the ideal value of K is 3. Following that, K-means clustering is used to divide the dataset into three groups. After obtaining the produced groups for every data point, a decision tree classifier was used to discover the pattern of the dataset's clustering. Following detection, Random Forest Tree was used to identify the most critical components in the dataset, which turned out to be the location and incident.

As text data was present in the dataset, text was preprocessed and converted to vectors of different dimensions for features. These dimensions were then reduced using PCA and visualised to find out string patterns between the features related to gender of the victim.

About 84% of the Bangladeshi women are continuously being sexually harassed in public places. It is practically impossible to find a woman who has not been sexually assaulted. Additionally, one out of ten men has been assaulted according to [16]. Our study takes a step forward towards making the street and the journey by the users safer. There are features in this system which will make the trip taken by the user a lot more secure and reliable. Factors like being able to see the "red zones" in the path and choosing path accordingly, alerting the user after entering the "red zone", the sharing location property after entering any "red zone" makes the users more cautious. The users will also be able to see the statistics of the sexual assault cases reported by the victims or the witnesses which will make the society more aware and alert about the rapid growth of the objectification of women and how the society is getting more and more unsafe by every passing day. The reporting feature will inspire more people to stand up for what is right even by staying anonymous.

Our proposed system helps to overcome a fair amount of challenges the society has been facing for decades. This system can be one sophisticated solution to many complications especially women have to face, everyday, while travelling.

6.2 Future Work

Our future work will involve learning more about graph machine learning and using it in our project to make it more meaningful and efficient. Even if the topic is fascinating, it would be more suited to investigate topics such as graph machine learning, density measurement, and so on.

Further information can be extracted from PCA. As PCA reduces the dimension without much of the information being lost, it clusters the highly correlated variables among the documents. These clusters can be used for Correlation Analysis to predict meaningful relationships between the variables. Eventually this can be used to anticipate semantic analysis of these variables. To a greater extent, NLP could also be used to detect the weight of the variables related to each. This could be done by feature extraction and feeding into machine learning algorithms.

As future work, the aim is to build a map which will portray the zones where sexual harassment cases occurred and the zones will be classified according to the clusters that were produced from the clustering algorithm used.

We also plan to take certain occasions into consideration when sexual harassment cases seem to increase by a large margin. Such occasions include Pohela Boishak, Ekushey February and many other occasions. The webapp would be able to give the statistics of the rise in sexual harassment cases during these occasions.

Bibliography

[1] W.-M. S. A. Famili, *Data preprocessing and intelligent data analysis*, Sep. 1996. DOI: 10.3233/IDA-1997-1102.

[2] S. Huda, *Sexual harassment and professional women in bangladesh*, Jan. 2003. DOI: https://doi.org/10.1163/1571815032120004..

[3] L. A. R. Kimberly Fairchild, *Everyday stranger harassment and women's objectification*, Sep. 2008. DOI: https://doi.org/10.1007/s11211-008-0073-0.

[4] V. J. C. R. Gentleman, *Unsupervised machine learning*, 2008. DOI: 10.1007/978-0-387-77240-010.

[5] B. P. Charles M. Judd, *Complete anonymity compromises the accuracy of self-reports*, Sep. 2011.

[6] "What is sexual harassment? in bangladesh how deep this problem is causing problem? is there any law according to bangladeshi justification? if there is any is it properly imposed or not? – explain illustrate.," Jul. 2011.

[7] *Red dot foundation, home page - safecity*, Dec. 2012.

[8] S. I. A. Syed Ishtiaque Ahmed, *Protibadi: A platform for fighting sexual harassment in urban bangladesh. session*, 2014. DOI: http://dx.doi.org/10.1145/2556288.2557376.

[9] C. Young, *Harassmap: Using crowdsourced data to map sexual harassment in egypt*, Mar. 2014.

[10] N. S. Grove, *The cartographic ambiguities of harassmap: Crowdmapping security and sexual violence in egypt*, 2015. DOI: https://doi.org/10.1177/0967010615583039.

[11] S. Kaushik, *An introduction to clustering and different methods of clustering*, Nov. 2016.

[12] A. A. Suzanne Goodney Lea Elsa D'Silva, *Women's strategies addressing sexual harassment and assault on public buses: An analysis of crowdsourced data*, 2017. DOI: I10.1057/s41300-017-0028-1.

[13] B. Bonaros, *K-means elbow method code for python*, Aug. 2019.

[14] T. Huda, "Sexual harassment and the law: Where's the problem?," Jun. 2019.

[15] F. Malik, *Machine learning hard vs soft clustering*, Jun. 2019.

[16] Z. Nasreen, *Rape of males: It's all about patriarchy*, Aug. 2019.

[17] B. Stecanella, "Understanding tf-id: A simple introduction," May 2019.

[18] A. Alif, "Male victims of rape suffer in silence," *Dhaka Tribune*, Oct. 2020.

[19] S. I. Sultana, “How the hierarchical clustering algorithm works,” Dec. 2020.

[20] *Unsupervised learning*, Sep. 2020.

[21] “84 percent of women facing sexual harassment,” *UNB*, Apr. 2021.

[22] R. Agrawal, “Must known techniques for text preprocessing in nlp,” Jun. 2021.

[23] T. Bock, *What is hierarchical clustering?* 2021.

[24] R. I. Sifat, *Sexual violence against women in bangladesh during the covid-19 pandemic*, Oct. 2022. DOI: 10.1016/j.ajp.2020.102455.

[25] A. L. Barstow, “Rape,”

[26] “Effects of sexual assault and rape,” *JOYFUL HEART FOUNDATION*,

[27] *K-means advantages and disadvantages.*

[28] S. B. R. Mohammed Eunus Ali, *Safestreet: Empowering women against street harassment using a privacy-aware location based application.*

[29] V. Powell, “Principal component analysis,”

[30] “What is sexual harassment?,”