**INTERNET DILATION AND IT’S EFFECT**



***We are all now connected by the Internet, like neurons in a giant brain.***

**~ Stephen Hawking**

**Acknowledgement.**

* We would like to express our heartfelt thanks and gratitude to all, who out of concern had come forward in completing our project.
* We would like to thank our Principal, Vice-Principal, HOD, Faculty members and the Non-Teaching staff members for providing us with the necessary logistics, infrastructure and the study environment every time we were in need of it.
* We would like to express our special word of thanks to our coordinator and mentor Mr. Chaitanya Alshi who has helped us, guided us and moreover being there to give us the right direction to conduct and complete our project with the right insights.
* We would also like to convey our heartfelt thanks to our HOD Mr. Amrit Rajwadkar for helping us in documenting the necessary permissions.
* We would also like to thank our teachers Mrs. Daksha Sawant, Mrs. Manjiri Vartak, Mr. Dhaval Shah, Mr. Krishna, Mr. Laxman Ms. Madhura Barve for helping us in collecting our data and also guiding and helping us in our analysis.
* And a huge thank you to each and every one who have helped us in every way possible right from the start of our project by means of pitching in their ideas, collection of data to the completion of our project.

**Introduction**

The internet has been one of the most transformative and fast-growing technologies. Day by day the number of internet users has been increasing drastically, and we as individual’s are becoming more and more dependent of this driving force of technology. The internet though has been a very strong instrument in shaping the culture of today’s modern world and played a major role in getting the world together and given each one of us a power way beyond our wildest imagination. But, alas, *“That which glitters is not always Gold”.*

The wide dependency on the Internet has also brought along the many problems, health issues, threats etc. unknown to mankind.

We as a group are trying to study the various problems and difficulties that we may tend to face due to excess use of the internet now or in the long run and to provide or take the necessary measures to overcome this issue.

**STEPS INVOLVED IN CONDUCTING THE SURVEY**

Step 1: Defining our objectives.

Step 2: Specifying information needs.

Step 3: Designing questionnaires.

Step 4: Pilot Survey.

Step 5: Modifying Questionnaires.

Step 6: Actual Survey.

Step 7: Data Cleaning.

Step 8: Coding of Data.

Step 9: Analysis of Data.

Step 10: Interpretation of Data.

Step 11: Conclusion.

**Collection of Data**

**Graphical Representation**

1. **Gender**
2. **Age**
3. **Type of family**
4. **Educational qualification**
5. Occupation

**FACTOR ANALYSIS**

**What is Factor Analysis?**

Factor analysis is a way to take a mass of data and shrinking it to a smaller data set that is more manageable and more understandable. Thus, it is called a ‘Dimension Reduction’ technique.

The key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) variable.

In every factor analysis, there are the same number of factors as there are variables.  Each factor captures a certain amount of the overall variance in the observed variables, and the factors are always listed in order of how much variation they explain. The factors that explain the least amount of variance are generally discarded.

**What is our objective?**

We wish to study the prime underlying factors behind when people use the Internet and how similar are these usage patterns for the variables.

**How do we go about this?**

We have performed Factor Analysis on SPSS for the ease of procedure it provides.

The audience was asked that on a scale from ‘Always’ to ‘Never’ how frequent is their use of Internet in the following situations.

There were 20 variables (You can refer the question at the end under the title questionnaire).

Factor Analysis can only be run if the variables are correlated with each other and the sample size needs to be adequate (enough in amount).

These two prime criteria are tested using **Bartlett’s Test of Sphericity and Kaiser-Meyer-Olkin test** respectively.

**Bartlett’s Test of Sphericity:** It tests whether the correlation matrix of the given variables is an Identity matrix or not.

It checks if you are permitted to use factor analysis as the given variables needed to be correlated to each other in order to recognise their underlying factor or reason.

To proceed further, your null hypothesis should be rejected implying that the correlation matrix is not an identity matrix and that correlations exist.

The following is the hypothesis:

Hₒ: Population correlation matrix is an identity matrix

H₁: Population Correlation matrix is not an identity matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **V1** | **V2** | **V3** | **V4** | **V5** | **V6** | **V7** | **V8** | **V9** | **V10** | **V11** | **V12** | **V13** | **V14** |
| **V1** | **1** | **0.378** | **0.082** | **0.361** | **0.216** | **0.24** | **0.011** | **0.045** | **0.167** | **-0.016** | **0.196** | **0.155** | **0.087** | **0.268** |
| **V2** | **0.378** | **1** | **0.245** | **0.348** | **0.296** | **0.516** | **0.239** | **0.167** | **0.225** | **0.14** | **0.22** | **0.219** | **0.311** | **0.242** |
| **V3** | **0.082** | **0.245** | **1** | **0.169** | **0.273** | **0.271** | **0.336** | **0.271** | **0.224** | **0.339** | **0.104** | **0.204** | **0.334** | **0.019** |
| **V4** | **0.361** | **0.348** | **0.169** | **1** | **0.223** | **0.257** | **0.152** | **0.196** | **0.237** | **0.146** | **0.216** | **0.227** | **0.174** | **0.181** |
| **V5** | **0.216** | **0.296** | **0.273** | **0.223** | **1** | **0.4** | **0.242** | **0.197** | **0.175** | **0.14** | **0.317** | **0.257** | **0.302** | **0.245** |
| **V6** | **0.24** | **0.516** | **0.271** | **0.257** | **0.4** | **1** | **0.276** | **0.231** | **0.199** | **0.188** | **0.225** | **0.244** | **0.354** | **0.221** |
| **V7** | **0.011** | **0.239** | **0.336** | **0.152** | **0.242** | **0.276** | **1** | **0.388** | **0.211** | **0.429** | **0.241** | **0.3** | **0.337** | **0.057** |
| **V8** | **0.045** | **0.167** | **0.271** | **0.196** | **0.197** | **0.231** | **0.388** | **1** | **0.424** | **0.413** | **0.243** | **0.295** | **0.28** | **0.13** |
| **V9** | **0.167** | **0.225** | **0.224** | **0.237** | **0.175** | **0.199** | **0.211** | **0.424** | **1** | **0.347** | **0.263** | **0.29** | **0.204** | **0.162** |
| **V10** | **-0.016** | **0.14** | **0.339** | **0.146** | **0.14** | **0.188** | **0.429** | **0.413** | **0.347** | **1** | **0.263** | **0.28** | **0.262** | **0.007** |
| **V11** | **0.196** | **0.22** | **0.104** | **0.216** | **0.317** | **0.225** | **0.241** | **0.243** | **0.263** | **0.263** | **1** | **0.488** | **0.215** | **0.234** |
| **V12** | **0.155** | **0.219** | **0.204** | **0.227** | **0.257** | **0.244** | **0.3** | **0.295** | **0.29** | **0.28** | **0.488** | **1** | **0.325** | **0.226** |
| **V13** | **0.087** | **0.311** | **0.334** | **0.174** | **0.302** | **0.354** | **0.337** | **0.28** | **0.204** | **0.262** | **0.215** | **0.215** | **1** | **0.235** |
| **V14** | **0.268** | **0.242** | **0.019** | **0.181** | **0.245** | **0.221** | **0.057** | **0.13** | **0.162** | **0.007** | **0.234** | **0.234** | **0.235** | **1** |

**The Correlation Matrix**

**The Correlation Matrix**

**KMO Test of Sampling Adequacy**: Rule of Thumb is that to get a decent output, the sample size should be:

The number of variables \* 20= Number of respondents (sample size)

So, we had 14 \*20 =280.

In our case, we had 733 respondents (after deletion of few variables to make the model better) which is a decent enough sample size to run factor analysis.

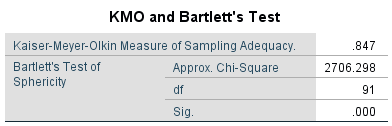
The rule for interpretation of the value returned by KMO which lies between 0 and 1 is as follows:

0.8 and higher: Great

0.6-0.7: Acceptable

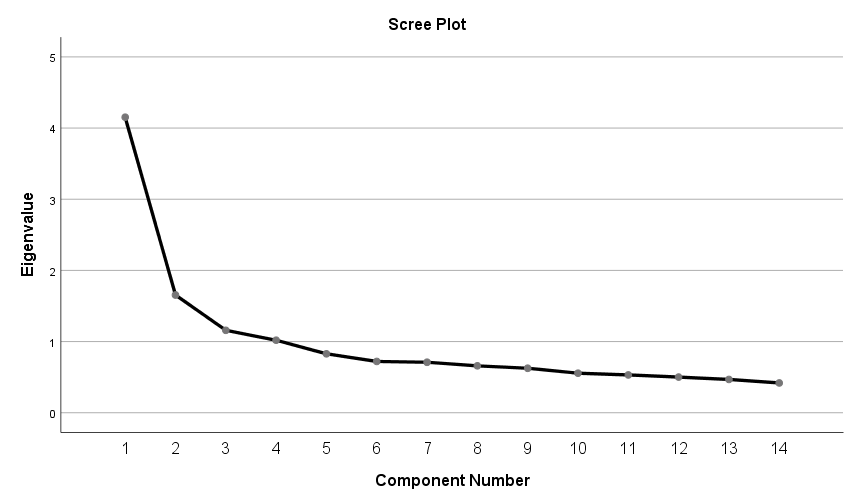
0.5 and lower: Not Acceptable

The results for both these tests can be seen below:



As you can see, our KMO gives us a value of 0.8 and higher and as you can see from the correlation matrix above our Bartlett’s test of Sphericity gives a significant result.

We go forward with factor extraction. For this we give a condition to the software to give us factors who have **eigen values > 1** and we obtain the scree plot as well.



As you can see after component 4 there is no significant variation and after component 4 all components have **eigen values <1.**

Thus, we know we will have 4 factors extracted.

**Principal Component Analysis Extraction Method:**

PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables.

It then removes this variance and seeks a second linear combination which explains the maximum proportion of the remaining variance, and so on.

Initially, we obtained a total Variance explained by 4 factors (as compared to 20 variables explaining 100% variance) as 51%.

We then checked which variables had low communalities.

**What are Communalities?**

This is the proportion of each variable’s variance that can be explained by the factors. If communalities for a variable are low (0.0-0.4) then that variable may struggle to load significantly on any factor.

We deleted variables having low communalities one by one and checked its effect on total variance explained by 4 factors.

Finally, we obtained a total variance of 57.030% with 14 variables.

|  |  |  |
| --- | --- | --- |
| **Communalities** | | |
|  | Initial | Extraction |
| [At home sitting idly] | 1.000 | .641 |
| [Need to reduce your mental stress] | 1.000 | .623 |
| [At the stadium to watch football, basketball etc.] | 1.000 | .543 |
| When you are….. [In bed about to sleep] | 1.000 | .559 |
| [At a place to repair your car, house appliances, etc] | 1.000 | .491 |
| [Facing emotional stress] | 1.000 | .595 |
| .. [Sitting in a religious place (e.g., church, mosque) and activities like sermon or prayer is yet to start] | 1.000 | .533 |
| [In the company of friends/family/colleagues having fun] | 1.000 | .554 |
| [At home Watching TV, news, football, films, sports, etc.] | 1.000 | .583 |
| [Go to the cinema house to watch movie(s)] | 1.000 | .607 |
| [A passenger in a car/bus/train for at least 2 min] | 1.000 | .642 |
| [Waiting for your boss/teacher in her office for at least 2 min ] | 1.000 | .604 |
| Facing financial challenges] | 1.000 | .526 |
| [Online school/ job-related work (project/homework)] | 1.000 | .483 |
| Extraction Method: Principal Component Analysis. | | |

We then went further in observing our component matrix.

To get more interpretable factors we applied Orthogonal Rotation so that the factors obtained after observing the rotated component matrix are uncorrelated. From the orthogonal types we chose the Varimax Rotation which preferably used.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rotated Component Matrixa** | | | | |
|  | Component | | | |
| 1 | 2 | 3 | 4 |
| [Go to the cinema house to watch movie(s)] | .754 | .183 | -.048 | .058 |
| [In the company of friends/family/colleagues having fun] | .708 | .133 | .092 | .164 |
| [At home Watching TV, news, football, films, sports, etc.] | .641 | -.062 | .361 | .197 |
| .. [Sitting in a religious place (e.g., church, mosque) and activities like sermon or prayer is yet to start] | .540 | .471 | -.109 | .093 |
| [Facing emotional stress] | .064 | .686 | .324 | .125 |
| Facing financial challenges] | .214 | .644 | -.035 | .252 |
| [At a place to repair your car, house appliances, etc] | .011 | .581 | .145 | .364 |
| [At the stadium to watch football, basketball etc.] | .430 | .566 | .061 | -.183 |
| [At home sitting idly] | -.080 | .075 | .773 | .175 |
| When you are….. [In bed about to sleep] | .233 | .097 | .699 | .083 |
| [Need to reduce your mental stress] | .054 | .539 | .570 | .068 |
| [A passenger in a car/bus/train for at least 2 min] | .265 | .082 | .093 | .746 |
| [Waiting for your boss/teacher in her office for at least 2 min ] | .354 | .172 | .037 | .669 |
| [Online school/ job-related work (project/homework)] | -.151 | .177 | .272 | .595 |
| Extraction Method: Principal Component Analysis.  Rotation Method: Varimax with Kaiser Normalization.a | | | | |
|  | | | | |

We checked which variables had the highest Factor loading under which Factor.

**What is a factor loading?**

It is nothing but the correlation coefficients between the variables (rows)and the factors (columns)

If a variable didn’t make sense getting loaded under a factor, we checked the second highest loading and checked if it made sense loading under that factor. If it did, we have gone ahead with it.

But we need to keep in mind that it can’t be done always and with many variables. Also, that we always check for second highest factor loading.

Therefore, we obtain the four situations where the usage pattern of the internet is similar across the respondents and we have named them as follows.

1.Entertainment Related Periods.

a. Go to the cinema house to watch movie

b. In the company of friends/family/colleagues having fun

c. At home Watching TV, news, football, films, sports, etc.

d. At the stadium to watch football, basketball etc.

2.Stress Related Periods.

a. Facing Emotional Stress

b. Facing Financial Challenges

c. Need to reduce mental stress.

3.Personal time Related Periods.

a. At home Sitting Idly

b. In bed about to sleep.

c. Online Job/School related homework.

4.Free Time Related Periods.

a. At a place to Repair your car/appliances.

b. A passenger in car/bus/train for at least 2 minutes.

c. Waiting for your boss/teacher in their office for at least 2 minutes

We conclude that the respondents use internet when they are in these phases of their day and that the usage pattern is similar for the variables that loaded under these factors.

**Association Rules**

In data mining Association rules are “if- then” statements that help us to show probability of relationships between data items within large data sets in various types of databases.

Association rule mining has number of applications, and are widely use to discover correlation in transactional data.

Association rules are very useful in various fields like Sales, Banking and Marketing. Therefore, companies like Walmart, D-Mart, Flipkart etc. use Association rules to analyze customer buying behavior.

**How does Association Rule Work:**

Association rule mining involves the use of machine learning models to analyze data for patterns or co-occurrences in a database. It identifies frequent “if-then” associations, which are called Association rules.

An association rule has two parts,

1. **Antecedent (if)**
2. **Consequent (then)**

Antecedent is an item found within the data and Consequent is an item found in combination with the antecedent.

Association rules is based on three criteria,

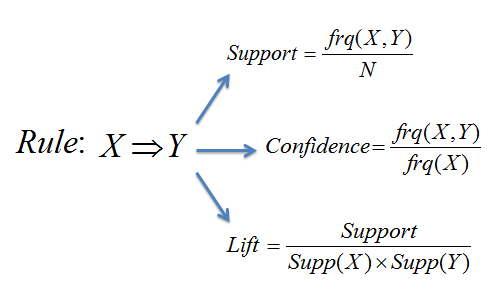
* **Support:** Support is an indication of how frequently an item appears in the dataset.
* **Confidence:** Confidence is the percentage value that shows how frequently the rule-head occurs among all the groups containing the rule-body.
* **Lift:** Lift summarizes the strength of association between the product of the left- and right-hand side of the rule; larger the lift the greater the link between the two products.

Lift > 1 positive association, Lift =1 no association, Lift < 1 negative association.

**Apriori Algorithm**

Apriori algorithm is classical algorithm in data mining. It is used for mining the frequent itemset and relevant association rules. It is very important for effective Market Basket Analysis and helps the customers in purchasing their items with more ease which increases the sales of markets.

Formula for rule,



Example:

Consider an association rule as follows for the following table,

{Baby-oil} => Diaper

|  |  |
| --- | --- |
| TID | ITEMS |
| 1 | Baby-oil, Diaper, Aspirin, Egg |
| 2 | Coffee, Aspirin, Milk |
| 3 | Diaper, Milk, Baby-oil, Aspirin |
| 4 | Biscuit, Coffee, Bread |
| 5 | Milk, Baby-oil, Bread, Diaper |
| 6 | Baby-oil, Aspirin |

For the above rule support, confidence and lift as follows;

**Support** = = = 0.5

**Confidence**  = = = 0.75

**Lift**  = = = 1.

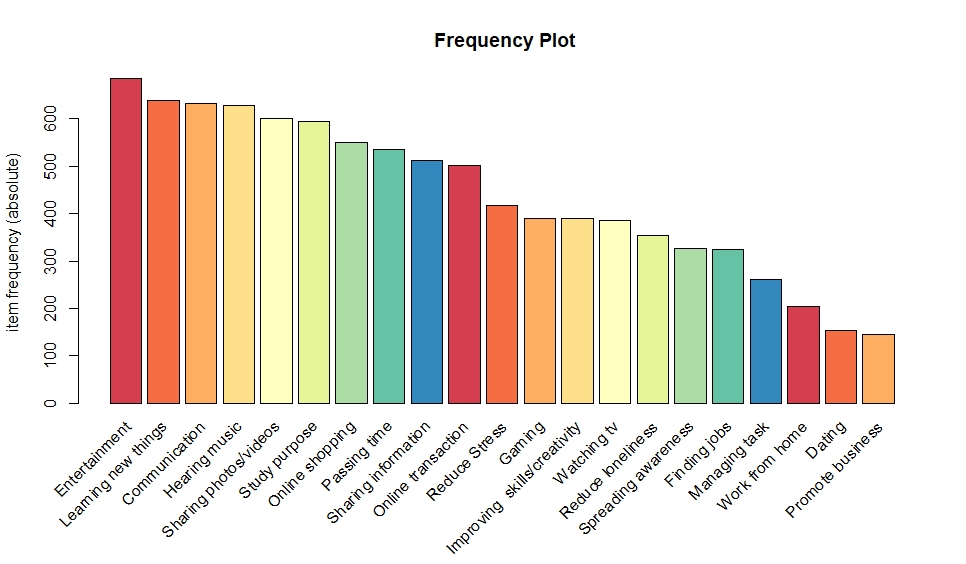
**Support** shows that given rules has found 50% of overall transaction.

**Confidence** indicates that 75% rule has been found true.

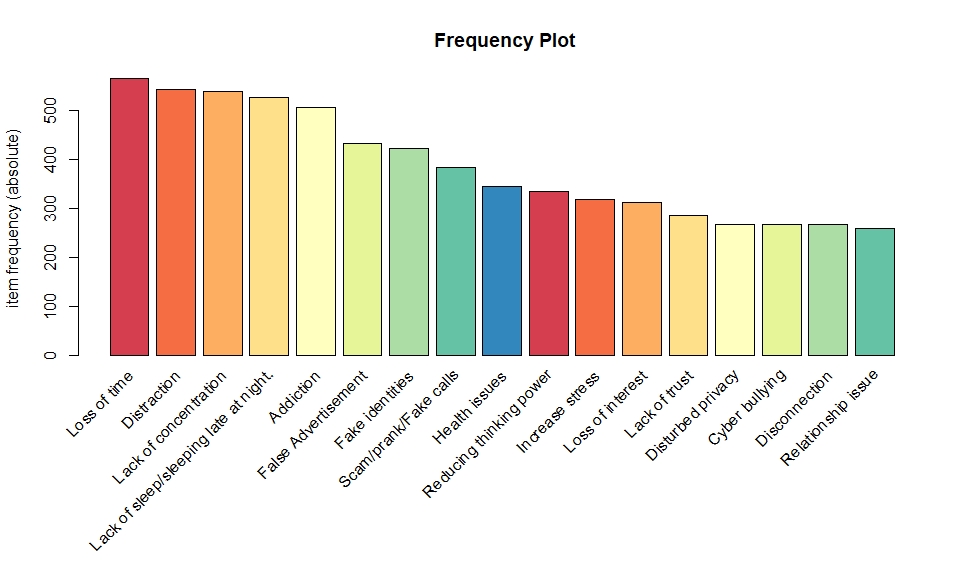
**Lift** value is 1.5 shows that {Baby-oil} has positive relation with {Diaper} .

**Objective: To study how negative instances associate to people’s activity on their daily usage of internet.**

Graphical representation of daily instances which people prefer by using internet,



Graphical representation of negative instances which people faced/heard by using internet,



* By using association rules, we are trying to achieve the negative effects of internet on the people by using internet daily.
* To find out relevant rule we used Apriori algorithm.
* In Apriori algorithm we have to set our minimum support and confidence, so then we can get relevant rule.
* In this case we set the minimum support as **0.1** and confidence as **0.7.** and so we get the relevant rules generated as follows,

**Association Rules**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rules** | **Support** | **Confidence** | **Lift** |
| **{Dating} => {Lack of concentration}** | **0.118005** | **0.7886179** | **1.216217** |
| **{Gaming} => {Addiction}** | **0.127737** | **0.7241379** | **1.242675** |
| **{Gaming} => {Distraction}** | **0.131387** | **0.7448276** | **1.138008** |
| **{Passing time} => {Distraction}** | **0.324818** | **0.7437326** | **1.136335** |
| **{Watching tv} => {Distraction}** | **0.273723** | **0.7785467** | **1.189527** |
| **{Watching tv} => {Lack concentration}** | **0.262774** | **0.7474048** | **1.152658** |
| **{Sharing photos/videos} => {Distraction}** | **0.43674** | **0.722334** | **1.10364** |
| **{Gaming, Hearing music} => {Addiction}** | **0.105839** | **0.7310924** | **1.25461** |
| **{Gaming, Passing time} => {Loss of time}** | **0.10219** | **0.7368421** | **1.272446** |
| **{Gaming, Passing time} => {Distraction}** | **0.109489** | **0.7894737** | **1.206222** |
| **{Online shopping, Online transaction} => {Addiction}** | **0.116788** | **0.7559055** | **1.297191** |
| **{Spreading awareness, Watching tv} => {Distraction}** | **0.155718** | **0.8152866** | **1.245661** |
| **{Entertainment,Gaming,Passing time} => {Loss of time}** | **0.100973** | **0.7410714** | **1.279749** |
| **{Entertainment, Online transaction,Watching tv} => {Distraction}** | **0.128954** | **0.8548387** | **1.306092** |
| **{Hearing music,Online shopping,Online transaction} => {Addiction}** | **0.104623** | **0.7478261** | **1.283326** |
| **{Hearing music,Online transaction,Sharing photos/videos} => {Addiction}** | **0.125304** | **0.7463768** | **1.280839** |

In below donut chart we can see people’s dependency on internet,

By above donut chart we observed that 76% people’s think that they are very dependent on internet and 24% people’s think that they are not dependent on internet.

**Results: -**

It has been observed that all the rules have been found to be significant.

The Lift value for all the rules is greater than “1”, which indicates that the occurrence of the rule-body has positive effect on the rule-head.

* If the people are having a tendency to play online games in their daily routine, there is a possibility that the people will get Addicted, Distracted, and also lose their precious time.
* If people are using online Dating apps for personal purposes, there is a possibility that the people may face lack of concentration in their day-to-day activities.
* Nowadays the people prefer to shop more online as compared to the traditional way of shopping which leads to shopping of unnecessary items, by which they are more prone to shop online and get addicted over the course of time.
* “Time is Money” and so in order to save time people have adapted ways to make monetary transactions easy this has resulted in the ease of making payments but may be a triggering factor to shop online easily and also to get distracted and addicted to this habit.
* Many people tend to use the Internet in their free time to rather than getting themselves involved in some recreational activity such people may lose their precious time and would not be able to concentrate on other main priorities throughout the day.
* From kids to the older generation and every one tend to use the internet as medium to entertain ourselves by means of watching T.V., hearing music via our earphones and headphones to refresh and relax ourselves overuse of this medium of entertainment will have adverse effects on our health which may lead to distraction and loss of time.
* Today’s young as well as old generation are expressing their thoughts and feelings via various Social Media Platforms in order to gain attention or to express their views on the various ongoing scenarios or to take stand on many socio-political issues unaware of the fact that they might get distracted, addicted and slowly even lack concentration in their daily routine.

By this point we have observed that, if people increase their use of internet in their daily routine then, they are more likely to face problems like,

* Addiction,
* Distraction,
* Loss of time,
* Lack of concentration.

To avoid these problems in future, people have to decrease their use of internet in day to day life to some extent.

### DECISION TREES

### **Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.**

**Decision tree are used when:**

**1. We have high dimensional data.**

**2.Intuitive representation that is easily understood by humans.**

**3.Learning and classification are simple and fast.**

**4.We have a good accuracy data.**

**There are two types of decision tree: -**

**Classification Tree: where the target variable is categorical and the tree is used to identify the ‘class’ within which a target variable would likely fall into.**

**Regression Tree: where the target variable is continuous and tree is use to predict its value.**



### ****Terminology related to Decision Trees:****

**ROOT Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets.**

**SPLITTING: It is a process of dividing a node into two or more sub-nodes.**

**Decision Node: When a sub-node splits into further sub-nodes, then it is called decision node.**

**Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node.**

**Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.**

**Branch / Sub-Tree: A sub section of entire tree is called branch or sub-tree**

**Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.**

There are 4 different types of tree growing methods:

1. Gini
2. Entropy
3. Chi square
4. Reduction in variance (used for continuous target data)

**Gini**: - Gini works with categorical target variable. It performs only binary splits. Higher the value of Gini higher the homogeneity. CART (Classification and Regression Trees) uses Gini method to create binary splits.

**Entropy: -** A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values. Entropy uses to calculate the homogeneity of a sample. If a sample is completely homogeneous then entropy is zero and if the sample is equally divided then it has entropy of one.

The lower the entropy the better split, the pure segmentation. The higher the entropy more impurity, more diversity in segment, the more uncertainty. Therefore, entropy should be less.

**Chi-square: -** It works with categorical target variable. It can perform two or more splits. Higher the value of chi-square higher the statistical significance of differences between sub-node and the parent node. It generates tree called CHAID (Chi-square Automatic Interaction Detector)

**Reduction in Variance: -** Reduction in variance is an algorithm used for continuous target variables.

**Pruning**: - It should reduce the size of decision tree by removing nodes (opposite of splitting) without reducing predictive accuracy. There are many techniques for the tree pruning.

Reduced error pruning: - One of the simplest forms of pruning is reduced error pruning. Pruning and subtree selection based on minimizing the error rate in the validation partition at each pruning step and then in the overall subtree sequence. This is usually based on the misclassification rate for a categorical response variable, but ASE can also be used.

Types of Data partition:

**Training Dataset**: - The sample data use to fit model.

**Testing Dataset**: - The sample of data use to provide unbiased evaluation of final fit model on Training dataset.

**Model information**

|  |  |
| --- | --- |
| Split criterion used | Entropy |
| Number of branches used | 2 |
| Maximum tree depth requested | 8 |
| Maximum tree depth achieved | 8 |
| Number of observations read | 853 |
| Number of observations used | 853 |
| Number of observations in training set | 682 |
| Number of observations in testing set | 171 |

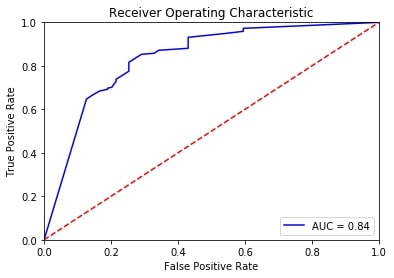
**Fit statistics for selected tree**

|  |  |  |
| --- | --- | --- |
|  | Training set | Testing set |
| Misclassification | 0.2196 | 0.2753 |
| Sensitivity | 0.8894 | 0.8303 |
| Specificity | 0.7075 | 0.7863 |
| Entropy | 0.7249 | 0.3663 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **Actual** | **Predicted**  **YES** | **Predicted**  **NO** | **Error Rate** | | Training | YES | 406 | 43 | 0.0957 | | NO | 92 | 112 | 0.4509 | | Testing | YES | 82 | 27 | 0.2477 | | NO | 30 | 32 | 0.4835 |   **Confusion Matrices** |

**Accuracy of Training data: 70.75%**

**Accuracy of Testing data: 78.63%**



AUC for testing dataset shows a good model fit measure.

**INTERPRETATION:**

- A person with education level (currently pursuing included) Graduate, Post Graduate and others support censorship.

-A person with education level (currently pursuing included) Graduate, Post Graduate and others having occupation Business or service with age above 25 years support censorship.

-A person with education level (currently pursuing included) Graduate, single having occupation service, age greater than 35 does not support censorship.

-A person educated up to HSC living in a Joint family supports censorship.

-A person educated up to SSC , having occupation student or service , living in a nuclear family does not support censorship.

-A person with education level up to SSC or upto HSC does not support censorship.

-A person with education level upto HSC , age less than 20 years does not support censorship.

**PARETO ANALYSIS**

Pareto Analysis is a statistical technique in decision making that is used for the selection of a limited number of tasks that produce significant overall effect. It uses the Pareto Principle. It is also known as the 80/20 rule. The idea is that by doing 20% of the work, you can generate 80% of the benefit of doing the whole job. This is also known as the "vital few" and the "trivial many" effect.

The Pareto principle has many applications in quality control. It is the basis of the Pareto diagram, one of the key tools used in total quality control and Six Sigma.

A Pareto chart is used to graphically summarize and display the relative importance of the differences between groups of data. A Pareto chart is used to graphically summarize the display information to show importance of various problems or causes of problems. It is essentially a special form of a vertical bar chart that puts items in order (from highest to lowest) relative to some measurable effect of interest such as frequency, cost or time.

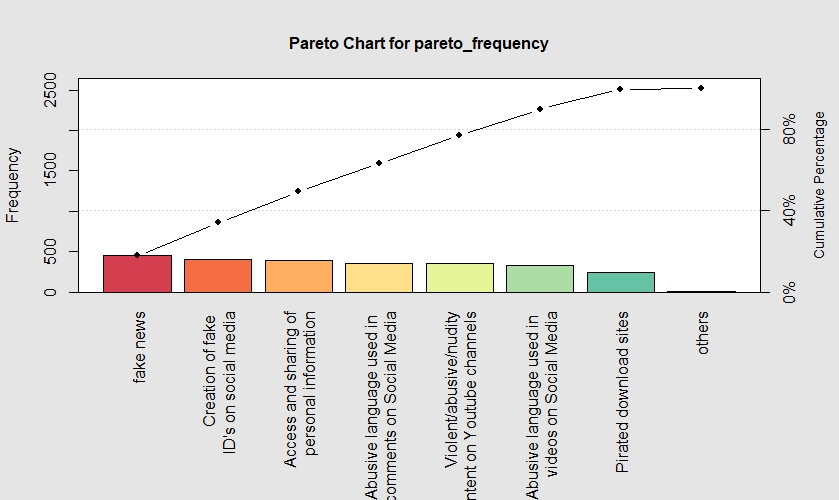
The chart is based on Pareto principle, which states that when several factors affect a situation, a few factors will account for most of the impact. The Pareto principle describes a phenomenon in which 80 percent of variation observed can be explained by mere 20 percent of the causes of that variation. The Pareto curve makes it clear as to where effort must be concentrated so as to give maximum effect.

The Pareto chart is a very simple but effective tool for prioritizing problem causes, which is why it is widely used for problem solving in the manufacturing industry .

The Pareto Chart is basically a descending bar graph that shows the frequencies of occurrences or relative sizes of various problems or causes of a particular problem.

The problem categories or causes are shown on x-axis of the bar graph. Aside from its main bar graph, the Pareto chart may also include a line graph that indicates the cumulative percentage of occurrences at each bar of the graph .This line graph referred to as the "cumulative percentage line", is used to determine which of the bars belong to the 'vital few ' and which ones are relegated to the 'trivial many'.

**Output using R**



Interpretation:

From Pareto analysis, we can see that the steps/platforms on which Government should apply censorship on the following platforms.

• Fake news;

• Creation of fake ID's on social media;

• Access and sharing of personal information;

• Abusive language used in comments on Social Media;

• Violent/abusive/nudity content on Youtube channels;

In reality it is not possible to curb all the problems on the Internet, but these are the major platforms where the government can apply censorship and so that the people in the long run will benefit from this.

**SENTIMENT/ TEXT ANALYSIS.**

**What is a Sentiment?**

It is a view or opinion that is held or expressed by an individual.

**What is Sentiment Analysis?**

Sentiment Analysis is a process of determining whether a piece of writing/content is positive, negative or neutral.

It is the most common text classification tool that analyses the income message and tells whether the underlying statement is negative, neutral or positive.

It is mostly used to analyse unstructured data i.e. data which is available in the form of text messages voice messages or even data which is not available in proper tabular or structured format etc.

The other word for sentiment analysis is opinion mining.

**What does it do?**

This type of analysis systematically identifies, extracts, quantifies (expresses) and studies the affect states (concepts that connect mental and physical process) of an individual.

**Where can it be applied?**

Sentiment Analysis can be applied to analyse customer service records, employee satisfaction levels by analysing their feedback forms, restaurant reviews, basically anything that where any person is trying to express his opinion or 8feelings.

**STEPS IN PERFORMING SENTIMENT ANALYSIS: -**

1. **Data Collection.**
2. **Forming a Corpus.**
3. **Pre-processing text.**
4. **Tokenization.**
5. **Forming a Term Document Matrix.**
6. **Visualization of the output.**

**EXPLANATION OF EACH OF TERMS:-**

1. **Data Collection.**

The data collection is the primary step before carrying out any analysis. There is no such compulsion regarding the collection of data for the analysis.

In our project the data was collected by means of a questionnaire (Online/Offline), where the respondents were told to express their opinion and views on “Internet, mobile phones are responsible for rapes” in the space provided in the questionnaire.

1. **Forming a Corpus.**

A Corpus is a collection of documents gathered at a single instance. Since we had only one column of entries, it was treated as a corpus i.e. a main body consisting of records from different users.

1. **Pre-processing text.**

Now the data required for analysis need to be made clean as it contains a lot of unnecessary information which maybe not required for the analysis. It is done in the following manner: -

1. Stemming: it is required to identify the root of the word from which it is formed.

For instance, *sailed, sailing, sailor* is stemmed to ***sail.***

1. Removing punctuation: It is required to remove the punctuations such as “**! @&<>, /” etc.**
2. Removing Stop word: Stop words are words which are used in sentences as connectors or conjunctions like ***the, is, that, for, which, etc.*** that don’t have any specific meaning.
3. Converting to lowercase: It is very necessary to convert the entire data into lowercase,

because the computer can take up to similar words ***ship*** and ***SHIP*** as two different words.

1. Replacing special words: like ***don’t*** to ***do not*** and ***wouldn’t*** too ***would not*** etc. in order to have a clearer context of the words used.
2. **Tokenization.**

It is the most important step which involves splitting of sentence into separate words. This is done in order to perform the text analysis that is the main step. For instance, let us consider the example: A sentence “Peace brings with it so many positive emotions that is worth aiming for in all circumstances” is converted into tokens as follows …

|  |  |  |  |
| --- | --- | --- | --- |
| *Peace* | *so* | *that* | *for* |
| *brings* | *many* | *is* | *in* |
| *with* | *positive* | *worth* | *all* |
| *It* | *emotions* | *aiming* | *circumstances* |

1. **Forming a Term Document Matrix.**

We then formulate the Term Document Matrix which is basically a matrix of words and their respective frequencies i.e. it displays in a tabular format the no. of times each word has occurred in a sentence, which can be further used in computing the Word Cloud and many much more analysis.

1. **Visualization of the output.**

Here we obtain the final output of our analysis and then we interpret the results representing it in the form of Graphs and Cloud.

**WORDCLOUD**

**What is a Word Cloud?**

* A Word Cloud is a popular visualization of words typically associated with text data. They are most commonly used to highlight popular or trending terms based on frequency of use of prominent and important. A Word Cloud is a beautiful, Informative image that communicates much in a single glance.

**What is the purpose of the Word Cloud?**

* The Word Cloud displays the most common words found in that text and shows them in a way that lets the viewer know what words are used in a text and with what kind of frequency.

**CLUSTER ANALYSIS.**

**Introduction:**

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group are more similar together than to those in other group.

The main task of Cluster Analysis is exploratory data mining and a common technique for statistical data analysis, it is used in many fields such as machine learning, Pattern recognition, Image analysis, Information retrieval, Bio-Informatics, Data Compression etc.

It is a measure of grouping records based on attributes that make them similar, if plotted the objects within the cluster will be closer together while the distance between the cluster will be far apart.

**Centroid Based Clustering:**

In centroid based clustering, a particular cluster is represented by a central vector, which may not be a member of a data.

**K-Means Clustering:**

K-means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups decided initially.

It classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e. they almost exhibit the same attribute within a particular cluster), whereas objects from different clusters are as dissimilar as possible (i.e. the distance between the cluster is as far as possible). In k-means clustering, each cluster is represented by its centre (i.e., centroid) which corresponds to the mean of points assigned to the cluster.

**Text Classification using K-means Clustering:**

Whenever a need arises when we need to retrieve the data, filtering the data and topic extraction, all of these have can be achieved by Text clustering also known as Document clustering.

To elaborate in a much more technical manner, Text Clustering is the application of cluster analysis to text-based documents. It uses NLP and machine learning to understand and categorize unstructured, textual data.

**How it works?**

Typically, the set of words that describe the topic matter are extracted from the document first.

Then then they are analysed for the frequency in which they are found in the document compared to other terms.

After which, Clusters of the set of words that describe the topic matter can be identified and the Auto-tagged.

**Objective:** To study the people’s opinion and views on ***“Internet, mobile phones are responsible for rapes”.***

The sentiment analysis was conducted with the help of two lexicons: -

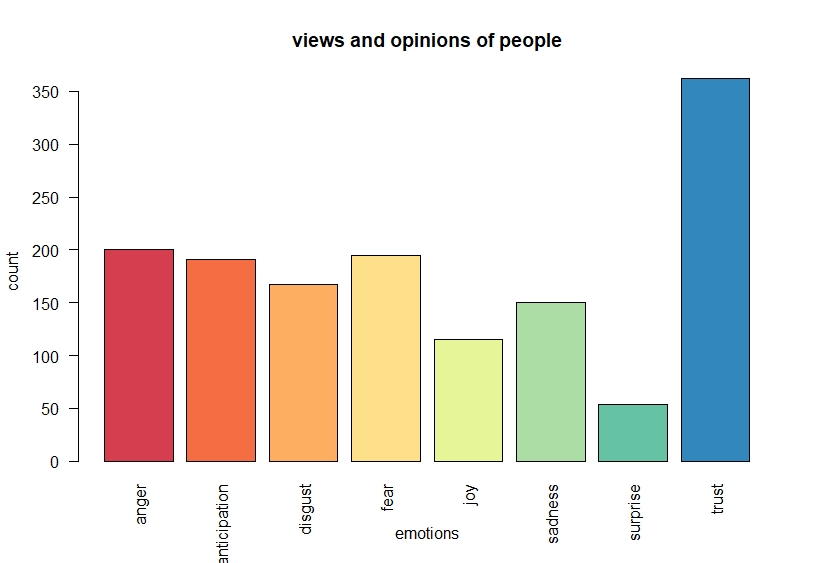
* **BING lexicon:** This lexicon categorizes words into a binary fashion. *(E.g. Positive and Negative).*
* **NRC lexicon:** This lexicon categorizes words into a multiple category fashion *(E.g. Anger,*

*Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust.*

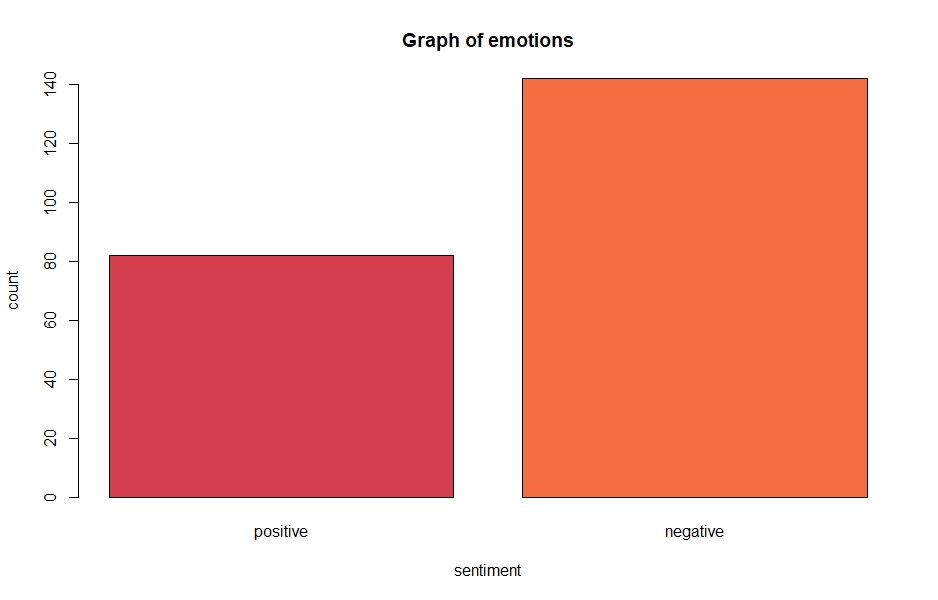
***(\* A lexicon is a vocabulary of words)***

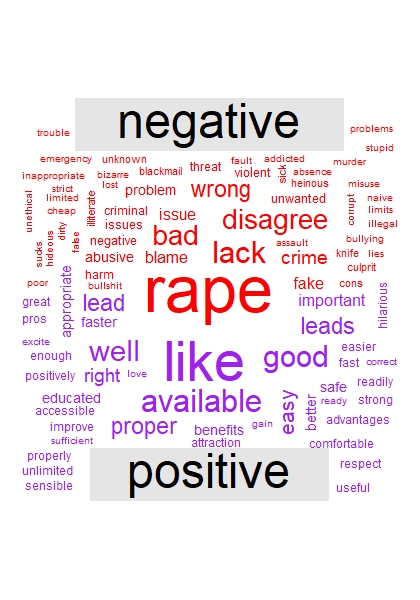
* ***Trust:*** *Firm Belief.*
* ***Anger****: Displeasure.*
* ***Fear****: Feeling of threat.*
* ***Anticipation****: Predict.*
* ***Disgust****: Strong Disapproval.*
* ***Sadness****: Unhappy.*
* ***Joy****: Feeling of happiness.*
* ***Surprise****: Unexpectedness.*

To accomplish this task, first the overall responses were collected, cleaned, pre-processed and visualized so as to understand what views were being expressed by the people on **“*Internet, Mobile Phones are responsible for rapes*”.**



* The above graph is obtained by NRC lexicon which showcases the emotions expressed in terms of counts classified in multiple categories.
* The responses obtained by the people expressed the emotion of ***“trust”*** to a greater extent followed ***“anger”,” fear”,” anticipation”, “ disgust”, “sadness”, “ joy”,” surprise”.***
* The people have strongly put forth their consent regarding the usage of Internet and Mobile phones and its after effects resulting in rape which can be seen in the graph corresonding to **“trust”**.
* The people have strongly expressed the emotions of **“Anger”, “Disgust”**, **“Fear”** and **“Sadness”** which are negative emotions, to a greater extent as compared to the emotion of **“Joy”** showing their consent that they disaprove that usage of Internet , Mobile phones results into Rapes.
* The emotion of **“Surprise”** is expressed to a vey lesser extent indicating that very few people were hesitant to express their view which contradicts with the emotion of **“Trust”**.
* The below graph is obtained **BING lexicon,** shows the emotions expressed in count classified in a Binary Fashion.

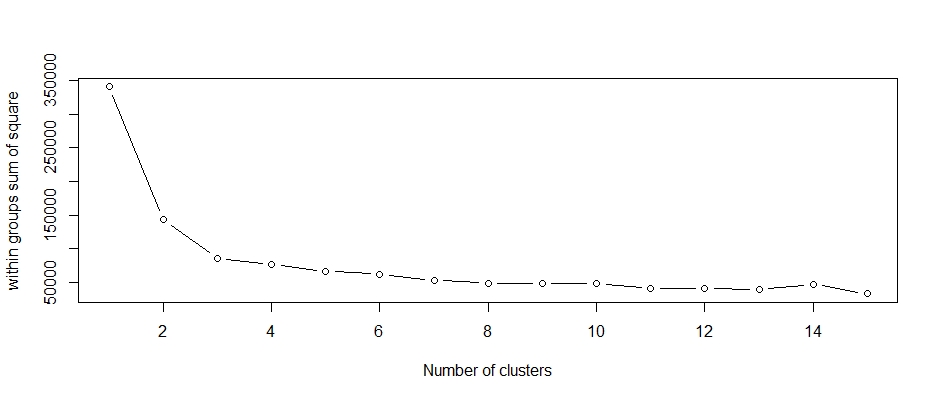




Now in order to gain more insights from what the people are trying to say, we tried to classify our data using cluster analysis by K-Means Clustering Method.

Now in order to determine the no. of clusters we used a Scree Plot which helped us to select the no. of clusters with respect to the Within Sum of Squares.

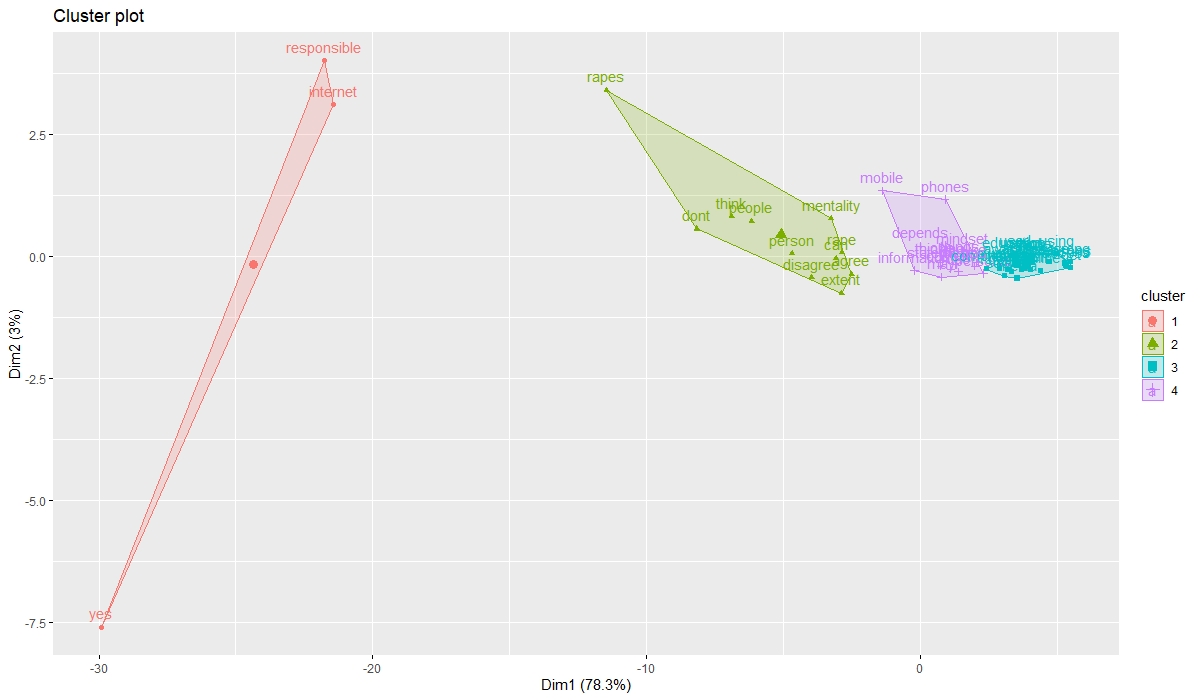
We obtained the following results.



By the Scree Plot the first 4 Clusters explaining Max within Sum of Square. And it gradually goes on decreasing.

So, we select only 4 Clusters. Now we proceed with our analysis by plotting the Cluster Plot.

**Cluster Plot:**



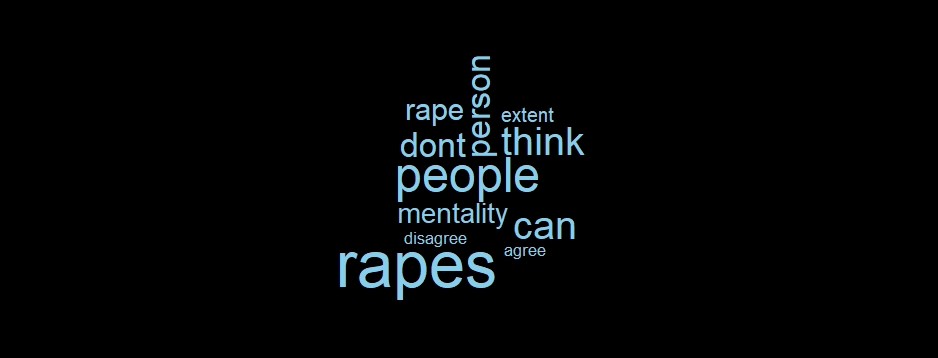
As you can see the following Clusters were obtained.

**Cluster Summary: -**

|  |  |
| --- | --- |
| **No of clusters=4** | |
| **Sr. No** | **No of Words in Each Cluster.** |
| 1 | 3 |
| 2 | 11 |
| 3 | 29 |
| 4 | 16 |

|  |  |
| --- | --- |
| **Within/Between cluster sum of squares by cluster:** | |
| **Corresponding to Cluster.** | **Sum of Squares** |
| 1 | 16566.88 |
| 2 | 28354.62 |
| 3 | 15718.94 |
| 4 | 15604.85 |
| **Between Sum of Squares**  264200 | |
| **Total Within Sum of Squares**  76245 | |
| **Total Sum of Squares**  340445 | |
| **Between\_SS / total\_SS**  0.7760 | |
| **Within\_SS / total\_SS**  0.2239 | |









**# R-studio codes for performing word cloud**

**getwd()**

**docs=read.csv("sentimentsNA.csv",header = F)**

**library(tm)**

**library(wordcloud)**

**library(RcolorBrewer)**

**corpus=Corpus(VectorSource(docs$V1))**

**corpus=tm\_map(corpus,removeNumbers)**

**corpus=tm\_map(corpus,removePunctuation)**

**corpus=tm\_map(corpus,tolower)**

**corpus=tm\_map(corpus,removeWords,stopwords('english'))**

**tdm=TermDocumentMatrix(corpus)**

**tdm=as.matrix(tdm)**

**tdm**

**w=sort(rowSums(tdm),decreasing = F)**

**set.seed(123)**

**wordcloud(words=names(w),freq = w,max.words = 100,random.order = T,min.freq = 0,color="green1",rot.per = 0.1)**

**# R-studio codes for performing sentiment analysis**

**library("lubridate") library(reshape2)**

**library(syuzhet) library(NLP)**

**library(ggplot2) library(tidyverse)**

**library(scales) library(stringr)**

**library(dplyr) library(tidyverse)**

**library(tidytext) library(dplyr)**

**library(wordcloud) library(sqldf)**

**docs=read.csv("sentimentNA.csv" ,header = F)**

**text=iconv(docs$V1)**

**sentiment=get\_nrc\_sentiment(text)**

**s=get\_nrc\_sentiment(text)**

**barplot(colSums(s),las=2,col=rainbow(10),ylab = "count",xlab = "emotions",main="views and opinions of people")**

**get\_sentiments('bing') %>%**

**filter(sentiment %in% c("positive","negative")) %>%**

**count(sentiment)**

**tokens=data\_frame(text=text)%>%**

**unnest\_tokens(word,text)**

**bing\_word\_count= tokens %>%**

**inner\_join(get\_sentiments("bing")) %>%**

**count(word,sentiment,sort=TRUE) %>%**

**ungroup()**

**sqldf("SELECT sentiment,COUNT(sentiment) as count FROM bing\_word\_count GROUP BY sentiment")**

**y=c("positive","negative")**

**x=c(82,142)**

**z=data.frame(**

**y=c("positive","negative"),**

**x=c(82,142))**

**barplot(x,names.arg=y,main="Graph of emotions",xlab="sentiment",ylab = "count",col=brewer.pal(8,"Spectral"))**

**tokens %>%**

**inner\_join(get\_sentiments("bing")) %>%**

**count(word,sentiment,sort=TRUE) %>%**

**acast(word ~ sentiment,value.var = "n",fill = 0) %>%**

**comparison.cloud(colors = c("red","purple"),max.words = 100)**

**# R-studio codes for text clustering**

**docs=read.csv("sentimentNA.csv" ,header = F)**

**library(tm) library(factoextra)**

**library(cluster) library(tidyverse)**

**corpus=Corpus(VectorSource(docs$V1))**

**corpus=tm\_map(corpus,removeNumbers)**

**corpus=tm\_map(corpus,removePunctuation)**

**corpus=tm\_map(corpus,tolower)**

**corpus=tm\_map(corpus,removeWords,stopwords('english'))**

**tdm=TermDocumentMatrix(corpus,control = list(minWordLength=c(1,Inf)))**

**t=removeSparseTerms(tdm,sparse=0.98)**

**m=as.matrix(t)**

**freq=rowSums(m)**

**freq=subset(freq,freq>1)**

**barplot(freq,las=2,col=rainbow(25))**

**distance=scale(m)**

**distance1=get\_dist(distance)**

**wss=0**

**for(i in 1:15){**

**kc=kmeans(distance1,centers = i,nstart = 20)**

**wss[i]=kc$tot.withinss**

**kc=kmeans(distance1,centers = 4)**

**str(kc)**

**fviz\_cluster(kc,data = distance1)**

**Software used: R-Studio,**

**# R-Studio Codes for performing association rule,**

**getwd()**

**list.files()**

**library(arules) # For generating rules**

**library(RColorBrewer) # For coloring the graph**

**dataset=read.transactions("last8.csv" ,sep=',',rm.duplicates = TRUE)**

**summary(dataset)**

**itemFrequencyPlot(dataset,topN=38,type="absolute",col=brewer.pal(9,'Spectral'),main=" Frequency Plot")**

**rules=apriori(data=dataset,parameter = list(support=0.1,confidence=0.7))**

**inspect(rules[1:100])**

**# R-Studio for pareto analysis**

**pareto= read.csv(file.choose())**

**View(pareto)**

**library(ggplot2)**

**library("qcc")**

**library("RColorBrewer")**

**pareto\_frequency=pareto$frequency**

**names(pareto\_frequency)=pareto$title**

**pareto.chart(pareto\_frequency)**

**library(ggplot2)**

**pareto.chart(pareto\_frequency, cumperc=seq(0,100 , by=40),col=brewer.pal(8,"Spectral"))**

**Software used :- python**

**python codes of decision tree**

**#loading the libraries**

**import os**

**import pandas as pd**

**import numpy as np**

**import matplotlib as plt**

**pip install graphviz**

**pip install pydot**

**#loading the dataset**

**os.chdir("C:\\Users\\91982\\Desktop")**

**df=pd.read\_csv("censorship.csv")**

**df.head()**

**#splitting the data in training and testing dataset**

**from sklearn.model\_selection import train\_test\_split**

**X = df.drop('dependent',axis=1)**

**y = df['dependent']**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.30)**

**#fitting a decision tree**

**from sklearn.tree import DecisionTreeClassifier**

**dtree = DecisionTreeClassifier(criterion = 'entropy')**

**dtree.fit(X\_train,y\_train)**

**predictions = dtree.predict(X\_test)**

**#calculating accuracy and confusion matrix**

**from sklearn.metrics import classification\_report,confusion\_matrix,accuracy\_score**

**conf\_matrix=confusion\_matrix(y\_test,predictions)**

**accuracy=accuracy\_score(y\_test,predictions)**

**conf\_matrix,accuracy**

**print(classification\_report(y\_test,predictions))**

**#visualizing a decision tree**

**from sklearn.externals.six import StringIO**

**from sklearn.tree import export\_graphviz**

**import matplotlib.pyplot as plt**

**import graphviz**

**from IPython.display import Image**

**from sklearn.externals.six import StringIO**

**from sklearn.tree import export\_graphviz**

**import pydotplus**

**features = list(df.columns[1:])**

**features**

**dot\_data = StringIO()**

**export\_graphviz(dtree, out\_file=dot\_data,feature\_names=features,filled=True,rounded=True)**

**graph = pydotplus.graph\_from\_dot\_data(dot\_data.getvalue())**

**Image(graph.create\_png())**

**#AUC curve**

**import sklearn.metrics as metrics**

**from sklearn.metrics import roc\_auc\_score**

**from sklearn.model\_selection import learning\_curve,GridSearchCV**

**probs = dtree.predict\_proba(X\_test)**

**preds = probs[:,1]**

**fpr, tpr, threshold = metrics.roc\_curve(y\_test, preds)**

**roc\_auc = metrics.auc(fpr, tpr)**

**param\_grid={'max\_depth': np.arange(3,10)}**

**gridS= GridSearchCV(dtree, param\_grid)**

**gridS.fit(X\_train,y\_train)**

**tree\_preds=gridS.predict\_proba(X\_test)[:,1]**

**tree\_performance = roc\_auc\_score(y\_test, tree\_preds)**

**print('DecisionTree:Area under the ROC curve={}'.format(tree\_performance))**

**HINDI QUESTIONAIRE**

लिंग:- O पुरुष O महिला O अन्य

आयु :\_\_\_\_\_\_\_\_\_\_\_\_

परिवार का प्रकार:- O पिता, माता या बच्चों के साथ O संयुक्त कुटुम्ब O अकेला

शिक्षा योग्यता (वर्तमान में समावेशी) :- O दसवीं तक O बारहवीं तक O ग्रेजुएट O पोस्ट ग्रेजुएट O अन्य

पेशा :- O विद्यार्थी O सेवा O व्यापार O घर की पत्नी/ पति O अन्य

आप निम्नलिखित स्थिति में कितनी बार इंटरनेट का उपयोग करते हैं? जब…….

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | कभी नहीँ | कभी कभी | शायद ही कभी | अक्सर | हमेशा |
| एक सेमिनार /वर्कशॉप /मीटिंग लेक्चर में| |  |  |  |  |  |
| घर पर व्यर्थ बैठे हैं| |  |  |  |  |  |
| अपने मानसिक तनाव को कम करने के लिए| |  |  |  |  |  |
| स्टेडियम में फुटबॉल, बास्केटबॉल और अन्य खेल देखने के लिए| |  |  |  |  |  |
| किसी के लिए प्रतीक्षा करना (जैसे दोस्त)/ या तो उनके घर में/ या किसी अनजान जगह पर| |  |  |  |  |  |
| संगीत, रेडियो, धार्मिक व्याख्यान सुनना| |  |  |  |  |  |
| सोने के समय में| |  |  |  |  |  |
| शैक्षणिक उद्देश्य के लिए एक पुस्तकालय में पढ़ना| |  |  |  |  |  |
| अपनी कार, घरेलू उपकरणों, आदि अन्य चीजों की मरम्मत करने के लिए एक जगह पर| |  |  |  |  |  |
| आप भावनात्मक तनाव का सामना कर रहे हैं| |  |  |  |  |  |
| शादी समारोह, जन्मदिन की पार्टी, रिसेप्शन, जैसे अन्य सामाजिक समारोहों में| |  |  |  |  |  |
| एक धार्मिक स्थान (जैसे, चर्च, मस्जिद) में बैठना और धर्मोपदेश या प्रार्थना जैसी गतिविधियाँ शुरू करना अभी बाकी है| |  |  |  |  |  |
| उन लोगों के बारे में और जानने की जरूरत है जिनसे आप बाहर मिलते हैं| |  |  |  |  |  |
| दोस्तों / परिवार / सहकर्मियों के साथ मस्ती करते हुए| |  |  |  |  |  |
| घर पर टीवी देखना, समाचार, फुटबॉल, फिल्में, खेल, अन्य चीजें| |  |  |  |  |  |
| आप सिनेमा घर में फिल्म देख रहे हो| |  |  |  |  |  |
| कम से कम 2 मिनट के लिए कार / बस / ट्रेन में प्रतीक्षा कर रहे यात्री के रूप में| |  |  |  |  |  |
| आप अपने कार्यालय में कम से कम 2 मिनट के लिए अपने बॉस / शिक्षक की प्रतीक्षा कर रहे हों| |  |  |  |  |  |
| पैसे की चुनौतियों का सामना करना| |  |  |  |  |  |
| ऑनलाइन स्कूल / नौकरी से संबंधित काम (परियोजना / होमवर्क)| |  |  |  |  |  |

आप दैनिक जीवन में इंटरनेट का उपयोग क्यों करते हैं? (आप कई विकल्प चुन सकते हैं)

|  |  |  |  |
| --- | --- | --- | --- |
| O जागरूकता फैलाना | O आप एक नई नौकरी खोज रहे हैं | O घर से काम | O वक़्त काटना |
| O व्यवसाय को बढ़ावा देना | O संवाद/संपर्क करना | O खेल | O टीवी देखना |
| O अकेलापन कम करना | O जानकारी फैलाना | O ऑनलाइन लेनदेन | O मनोरंजन |
| O तनाव कम करना | O एक से अधिक कार्य को संभालनाअध्ययन का उद्देश्य | O ऑनलाइन खरीदारी | O फ़ोटो / वीडियो साझा  करना |
| O नई चीज़ें सीखना | O कौशल / रचनात्मकता में सुधार | O संगीत सुनना | O डेटिंग |

क्या आपने इंटरनेट का उपयोग करके इस प्रकार की समस्याओं का सामना किया / सुना है? (आप कई विकल्प चुन सकते हैं)

|  |  |  |  |
| --- | --- | --- | --- |
| O ध्यान की कमी | O रुचि खोना | O लत | O झूठा विज्ञापन |
| O ध्यान भंग | O समय का नुकसान | O दरार | O स्वास्थ्य समस्या |
| O बढ़ता तनाव | O रिश्ते के मुद्दे | O साइबर बदमाशी |  |
| O नींद की कमी / रात को देर से सोना | O गोपनीयता भंग | O स्कैम / प्रैंक / फेक कॉल |  |
| O विचार शक्ति को कम करना | O भरोसा की कमी | O नकली पहचान |  |

क्या आपको लगता है कि आप इंटरनेट पर निर्भर हैं?

* हाँ |
* नहीं |

गोपनीयता की चिंता और जोखिम के बारे में आपकी क्या राय है?

* मुझे जोखिम के बारे में पता है|
* मुझे जोखिम के बारे में जानकारी नहीं है|
* मैं जोखिम के बारे में जागरूक और चिंतित हूं|
* मैं जागरूक हूं लेकिन जोखिम के बारे में चिंतित नहीं हूं|

निम्नलिखित के बारे में आपके क्या विचार हैं? यह मुझे परेशान करता है जब ……

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | पूरी तरह से सहमत | सहमत | न इस बात से सहमत है और न ही असहमत | असहमत | पूरी तरह से असहमत |
| सोशल मीडिया प्लेटफॉर्म या साइट मुझसे व्यक्तिगत जानकारी प्रदान करने के लिए कहते हैं| |  |  |  |  |  |
| विभिन्न ऑनलाइन प्लेटफ़ॉर्म मेरी व्यक्तिगत जानकारी को अपने डेटाबेस में संग्रहीत / जमा करते है, या बेचते हैं| |  |  |  |  |  |
| मेरी संग्रहीत/ जमा व्यक्तिगत जानकारी सुरक्षित नहीं है। |  |  |  |  |  |
|
| मेरी संग्रहीत/ जमा व्यक्तिगत जानकारी का उपयोग अनधिकृत कर्मियों द्वारा किया जाता है। |  |  |  |  |  |
| मुझे अपनी जानकारी पर नियंत्रण नहीं है। |  |  |  |  |  |
| मैं कई ऑनलाइन प्लेटफ़ॉर्म की गोपनीयता नीतियों को नहीं समझ पा रहा हूं। |  |  |  |  |  |

आपके अनुसार भारत में इंटरनेट को सेंसर(निरीक्षक) किया जाना चाहिए?

* हाँ
* नहीं

इंटरनेट पर किन प्लेटफॉर्मों पर सरकार को सेंसरशिप((निरीक्षक) लागू करनी चाहिए?

|  |  |
| --- | --- |
| O व्यक्तिगत जानकारी भेजना | O सोशल मीडिया पर फर्जी आईडी का निर्माण |
| O फेक न्यूज | O यूट्यूब चैनलों पर हिंसक / अपमानजनक / नग्नता सामग्री |
| O सोशल मीडिया पर टिप्पणियों में अभद्र भाषा का इस्तेमाल किया गया | Oअनधिकृत डाउनलोडिंग साइटें |
| O सोशल मीडिया पर वीडियो में अभद्र भाषा का प्रयोग किया गया | Oअन्य:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

"इंटरनेट, मोबाइल फोन रेप्स के लिए जिम्मेदार हैं" इस कथन/ बयान के बारे में आपकी क्या राय है?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |
| --- |
| **Internet Intertwining The World** |
| Q1) Gender : O Male O Female O Other |
| Q2) Age : \_\_\_\_\_\_\_\_\_\_ |
| Q3) Type of Family :- O Nuclear O Joint O Single |
| Q4) Educational Qualification (currently pursuing inclusive) |
| O Upto SSC O Upto HSC O Graduation O Post graduation O Other |
| Q5) Occupation:- |
| O Student O Service O Business O House Wife/Husband |
| O Other(specify):\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Q6) How often do you use the Internet in the following situation? |  |  |  |  |  | |  |  |  |  |  |  | | When you are at….... |  |  |  |  |  | |  |  |  |  |  |  | |  | **Never** | **Rarely** | **Sometimes** | **Often** | **Always** | | *At a seminar/workshop /meeting/lecture* |  |  |  |  |  | | *At home sitting idly* |  |  |  |  |  | | *Need to reduce your mental stress* |  |  |  |  |  | | *At the stadium to watch football, basketball etc.* |  |  |  |  |  | | *Waiting for someone (e.g., friends) either in their house or at a prearranged place* |  |  |  |  |  | | *Listening to music, radio, religious lectures etc* |  |  |  |  |  | | *In bed about to sleep* |  |  |  |  |  | | *Reading in the library for academic purpose e.g., recommended text for class* |  |  |  |  |  | | *At a place to repair your car, house appliances, etc* |  |  |  |  |  | | *Facing emotional stress* |  |  |  |  |  | | *At a social gathering like wedding ceremony, birthday party, reception etc.* |  |  |  |  |  | | *Sitting in a religious place (e.g., church, mosque) and activities like sermon or prayer is yet to start.* |  |  |  |  |  | | *Need to find out more about people you met offline* |  |  |  |  |  | | *In the company of friends/family/colleagues having fun* |  |  |  |  |  | | *At home Watching TV, news, football, films, sports, etc.* |  |  |  |  |  | | *Go to the cinema house to watch movie(s)* |  |  |  |  |  | | *A passenger in a car/bus/train for at least 2 min* |  |  |  |  |  | | *Waiting for your boss/teacher in her office for at least 2 min* |  |  |  |  |  | | *Facing financial challenges* |  |  |  |  |  | | *Online school/ job-related work (project/homework)* |  |  |  |  |  |   Q7) Why do you use the Internet Daily (You can tick multiple options)   |  |  |  |  | | --- | --- | --- | --- | | * Spreading awareness | * Finding jobs | * Dating | * Hearing music | | * Promote business | * Communication | * Work from home | * Passing time | | * Reduce loneliness | * Sharing information | * Gaming | * Watching tv | | * Reduce Stress | * Managing task | * Online transaction | * Entertainment | | * Learning new things | * Study purpose | * Online shopping | * Share photos/videos | |
|  |

Q8) Have you faced/heard these types of problems by using internet? (you can tick multiple options)

|  |  |  |  |
| --- | --- | --- | --- |
| * Lack of concentration | * Loss of interest | * Addiction | * Lack of sleep / sleeping late at night. |
| * Distraction | * Loss of time | * Disconnection | * Scam/prank/Fake calls |
| * Increase stress | * Relationship issue | * Cyber bullying |  |
| * False Advertisement | * Disturbed privacy | * Health issues |  |
| * Reducing thinking power | * Lack of trust | * Fake identities |  |

Q9) Do you think you are dependent on the internet? \* (Mark only one oval.)

* Yes
* No

Q10) What is your opinion about privacy concern and risk? \* (Mark only one oval.)

* I am aware about the risk
* I am not aware about the risk
* I am aware and concerned about the risk
* I am aware but not concerned about the risk

Q11) What are your views about the following? It bothers me when……………. (Tick any one option per row)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Strongly Agree | Agree | Neutral | Disagree | Strongly Disagree |
| Social media platforms or sites ask me to provide personal information. |  |  |  |  |  |
| Various online platforms store or sell or share my personal information in their database. |  |  |  |  |  |
| My stored personal information is not secure. |  |  |  |  |  |
| My stored personal information is accessible by unauthorized personnel. |  |  |  |  |  |
| I don’t have control over my own information. |  |  |  |  |  |
| I am not able to understand the privacy policies of many online platforms. |  |  |  |  |  |

Q12) According to you should internet be censored in India? \* (Mark only one oval.)

* Yes ( If Yes , attempt Q)13 and Q)14)
* No (if No, don’t attempt Q)13 go directly Q)14)

Q13) What platforms on the internet should the government apply censorship on? \* (Check all that apply.)

|  |  |
| --- | --- |
| * Access and sharing of personal information | * Creation of fake ID's on social media |
| * Fake news | * Violent/abusive/nudity content on Youtube channels |
| * Abusive language used in comments on Social Media | * Pirated download sites |
| * Abusive language used in videos on Social Media | * Other(specify) : - \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

Q) "Internet, mobile phones responsible for Rapes" what are your opinions about the same ?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Bibliography**

*References For research on the topic:*

<https://www.dovepress.com/relationships-between-internet-addiction-and-clinicodemographic-and-be-peer-reviewed-fulltext-article-NDT>

<https://www.surveymonkey.com/r/NZP7J8G>

<https://www.livemint.com/news/india/social-media-regulations-to-be-ready-by-january-2020-centre-to-supreme-court-11571664731405.html>

<https://www.debate.org/opinions/is-there-a-need-to-censor-social-media-in-our-society>

<https://evolvetreatment.com/blog/parents-think-teen-addicted-phone/>

<https://www.omicsonline.org/open-access/impact-of-internet-on-social-connections-in-family-system-a-surveystudy-of-residents-in-lahore-2151-6200-1000270.php?aid=89875>

*Factor Analysis :*

<https://youtu.be/5WQ9021kkEI>

<https://youtu.be/xPyklfpiTzs>

<https://youtu.be/j5q1E4_wq3k>

<https://youtu.be/d0TZreR_yTY>

<https://youtu.be/UgrCLMHDMb4>

<https://www.theanalysisfactor.com/factor-analysis-1-introduction/>

<https://www.statisticshowto.com/factor-analysis/>

<https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/>

<http://statwiki.kolobkreations.com/index.php?title=Exploratory_Factor_Analysis>

<https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/idh_fact_rot.html>

<https://www.researchgate.net/post/Factor_Analysis_Which_method_and_rotation_should_i_use>

*Association Rules:*

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

<https://www.udemy.com/?utm_source=adwords-brand&utm_medium=udemyads&utm_campaign=Brand-Udemy_la.EN_cc.INDIA&utm_term=_._ag_78279294239_._ad_387461093429_._de_m_._dm__._pl__._ti_kwd-310556426868_._li_1007785_._pd__._&utm_term=_._pd__._kw_udemy_._&matchtype=e&gclid=EAIaIQobChMI58b8havP6AIVTIiPCh1uhwzFEAAYASAAEgKSDPD_BwE>

<https://en.wikipedia.org/wiki/Association_rule_learning>

*Pareto Analysis and Decision Trees:*

<https://www.projectsmart.co.uk/pareto-analysis-step-by-step.php>

<https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952>

<https://towardsdatascience.com/decision-tree-in-python-b433ae57fb93>

<https://www.kdnuggets.com/2017/05/simplifying-decision-tree-interpretation-decision-rules-python.html>

<https://youtu.be/eY6KCsArFdw>

*Sentiment Analysis:*

<https://blog.bitext.com/what-is-sentiment-analysis-and-how-does-it-work>

<https://insidebigdata.com/2018/07/26/what-is-text-clustering/>

<https://reason.com/2017/07/03/the-true-link-between-porn-and-rape/?amp>

<https://www.indiatoday.in/technology/features/story/many-porn-sites-are-banned-on-jio-airtel-in-india-is-it-legal-for-people-to-access-them-despite-ban-1408033-2018-12-12>