Internet Censorship – A Data Driven Investigation

Gnanitha Garikipati & Jin Zhang

December 9, 2024

1. Introduction

1.1 Problem

Internet censorship restricts access to information, suppresses freedom of speech and controls the dissemination of online content across many regions worldwide. These censorship efforts are enforced through various technical methods such as DNS blocking, TCP/IP filtering and HTTP throttling. Analysing data sets that capture censored network traffic overtime allows us to identify the types of content being targeted, the frequency of censorship and how these practices evolve. Understanding these patterns is crucial for identifying the mechanisms used to suppress online information and revealing the broader implications for digital freedom.

1.2 Importance

Internet censorship undermines the free flow of information and human rights, particularly freedom of expression and access to knowledge. By analysing this dataset, we can uncover the techniques and trends behind censorship, providing transparency and evidence of how such practices evolve. This analysis is essential for developing tools to circumvent censorship, informing policy decisions end raising awareness about digital rights violations.

1.3 Impact

If we can accurately identify and understand censorship patterns and methods, it could lead to more effective anti-censorship tools and strategies to protect online freedom, what type of content is being censored and statistics of the censored events. Researchers, activists and policymakers could use these insights to advocate for digital rights and challenge oppressive censorship practices. Ultimately solving this issue would promote a more open internet ensuring equitable access to information and protecting freedom of speech.

This project focuses on analysing a censored dataset from Russia spanning October 2022 to identify censorship patterns and track their evolution over time. By examining if the data is censored, and what type of censorship such as DNS blocking, TCP/IP filtering or HTTP throttling. This analysis will highlight differences in censorship practices and how network patterns change during such events, providing insights into the enforcement techniques used and the dynamics of information suppression.

2. Data

2.1 Data Collection

For this project, we are using data sets provided by Professor Abdullah Mueen, originally sourced from the Open Observatory of Network Interference (OONI). OONI Is an open-source initiative committed to detecting the documenting instances of Internet censorship, traffic manipulation and network performance issues worldwide. Through various tests on thousands of websites, OONI collects measurements dot offer a transparent view of how censorship manifests across different regions and networks. Specifically, we are utilising a data set from October 2022 which contains over 1,000,000 instances. This data set includes cases where URLs

were found to be entirely inaccessible indicating total censorship. A sample entry from a data set illustrates some of the attributes:

Table 1: Censored Dataset

input(url)	http://www.weedy.be/
measurement_start_time(UTC)	01-10-2022 00:00
probe_asn	AS44020
probe_network_name	Modern Solutions LTD
resolver_asn	AS15169
get_status_tcp_ip_connection	{'217.70.184.50:443': {'status': False, 'failure': 'connection_refused_error'}, '217.70.184.50:80': {'status': True, 'failure': None}}
get_http_request_status_code	200
ooni_logic_http_blocked_or_not	HTTP_Unknown
http_blocked_or_not	HTTP_Blocked
blocking_type	dns

2.2 Data Preprocessing

Data preprocessing is critical to ensure high-quality inputs for data mining tasks, The following techniques were applied for the dataset:

1. **Mapping:** Organized all the attributes by mapping them to the relevant information they convey, indicating whether each attribute is completely filled, partially filled or blank. This mapping allows for a clearer understanding of the dataset's structure and access the dataset easily.
2. **Filtering the columns:** To prepare the dataset for training the model, we removed columns that provided direct indications of censorship presence. Out of the original 40 columns, we retained only 8 columns that were most relevant for identifying censorship patterns without explicitly signalling censorship. This preprocessing step ensures the model learns to detect censorship based on subtle patterns in the data rather than relying on obvious indicators. The final columns are:

Table 2: Columns in final dataset

Column Name	Description
input(url)	The URL being tested for censorship.
measurement_start_time(UTC)	The exact timestamp when the measurement began.
probe_asn	The Autonomous System Number of the probe (the system or network conducting the test)
probe_network_name	The name of the network to which probe belongs to.
resolver_asn	The Autonomous System Number (ASN) of the resolver (the system or network conducting the test).
resolver_ip	The IP address of the DNS resolver being used to resolve domain names during the measurement.
resolver_network_name	The name of the network or ISP associated with the DNS resolver IP address.
blocking_type	Specifies the type of blocking observed (e.g., DNS blocking, IP blocking, content filtering, etc.). This is used for validating the results.

3. **Handling Missing Data:** Removed some irrelevant attributes from the dataset, specifically those are blank or contained over 70% missing data in a row. This includes attributes like resolver_asn, resolver_ip, resolver_network_name, and blocking_type were blank in many rows. Removed all these rows from the dataset.
4. **Outlier Detection and Removal:** The attribute measurement_start_time has the data that belongs to the previous month, filtered out this data and removed the irrelevant data from the dataset. Checked for duplicates and removed those records.

Finally, after removing all these rows we are left with 970,000 rows, from which we extracted first 100,000 rows for training. The final dataset has 4 different classes in blocking_type, dns, tcp-ip, http and False. The first 3 indicates the presence of censorship and the False indicates the absence of censorship. However, the there is huge variation in the count of each class based on which the accuracy was measured finally. Figure 1 describes the statistics of classes.

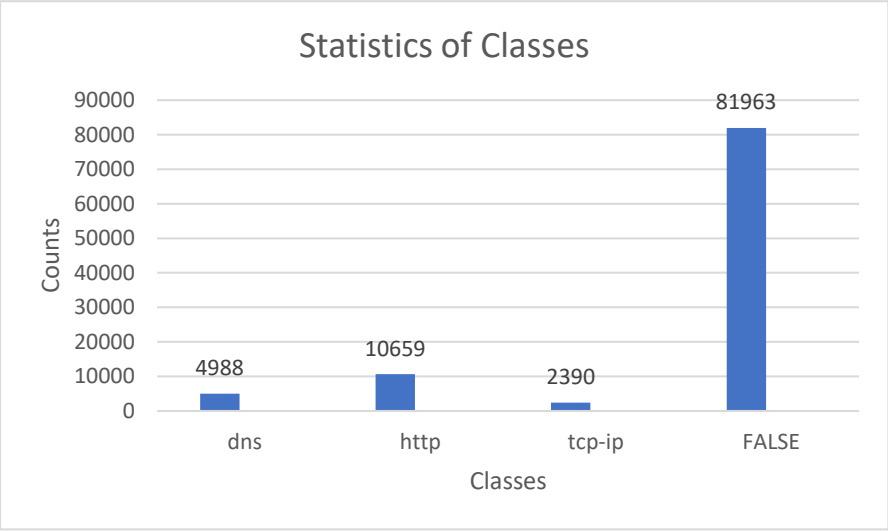


Figure 1: Statistics of class counts

3. Algorithm

The core objective of this project is to classify the data to determine whether censorship is present using a decision tree classification technique. The classification model's accuracy is assessed through 5-fold cross validation. If censorship is detected in the above step the data is further processed using K-means clustering technique to group, it based on specific features. The resulting clusters are evaluated for accuracy using the Rand index to ensure meaningful grouping.

To enhance the model performance and observe the variations in accuracy, the dataset is further pre-processed as follows:

1. **Extract Domain Names:** From input(url), striped all the text except the domain names.
2. **Combining Domain Names with Probe Information:** Created a new column by combining columns input(url), probe_asn and probe_network_name and named it as Domain_Probe.
3. **Combining Resolver Information:** Merge resolver_asn and resolver_ip column to resolver_asn_ip.
4. **Final columns:** Keep only the following columns, Domain_Probe, measurement_start_time (UTC), resolver_asn_ip, resolver_network_name, blocking_type

3.1 Classification using Decision Tree

Classes Conversion: For classification all the class labels are converted to 2 classes true or false, for cross-validation.

Classification Algorithm:

1) Load and Preprocess the Data
   a) Load the CSV file df_classification.csv using Python's csv library.
   b) Process each row:
      i) Convert numerical values with decimals to floats.
      ii) Convert integer-like values to integers.
      iii) Normalize non-numeric values (e.g., class labels) to lowercase strings.
   c) Store the processed rows into a data list for further processing
2) Cross-Validation
   a) Perform 5-fold Cross-Validation:
      i) Split the dataset into 5 equal parts (folds).
      ii) For each fold:
      iii) Use 4 folds for training and 1 fold for testing.
   b) Build the decision tree using the training data.
      i) Initialize an empty tree and a list of attributes to be considered.
      ii) Iterate until no attributes are left:
         (1) For each attribute, calculate its Gini Index using the subsets formed by splitting on that attribute.

$$\text{Gini Index} = 1 - \sum(\text{probability of each feature})^2$$

         (2) Select the attribute with the lowest Gini Index as the best splitting criterion.
         (3) Remove the selected attribute from the list of available attributes.
         (4) Add the selected attribute as a node in the decision tree.
      iii) For each value of the selected attribute:
         (1) If all data points in the subset belong to the same class, label the branch with that class.
         (2) Otherwise, retain the subset for further splitting
   c) Predict the class of each sample in the test set.
   d) Track the actual and predicted classes for each sample.
   e) Calculate the accuracy for the current fold as the ratio of correct predictions to the total number of test samples.
   f) Store the accuracy for each fold and compute the mean accuracy.
3) Calculate the overall confidence as the mean accuracy across all folds.

3.2 Clustering using K-means Clustering

For the clustering process, the text data is converted into numerical or binary format using the Word2Vec encoding algorithm. Before applying this encoding, the measurement_start_time(UTC) column was formatted to a standard datetime format. From this, a new column named day was created to represent the day of the month. The time of day was categorized into four quarters:

Q1: Midnight to 6 AM (00:00–06:00)

Q2: 6 AM to Noon (06:00–12:00)

Q3: Noon to 6 PM (12:00–18:00)

Q4: 6 PM to Midnight (18:00–00:00)

A new column named quarter was added to store these categories, and the original timestamp column was removed. The final set of columns includes Domain_Probe, day, quarter, resolver_asn_ip, resolver_network_name, and blocking_type.

Clustering Algorithm

1) The pre-processed dataset is loaded for clustering.
2) The feature values are standardized using the z-normalization technique.
3) To determine the optimal number of clusters, 2 methods are used.
   a) Elbow Method: The Elbow Method is used to find the optimal number of clusters by calculating the **Within-Cluster Sum of Squares (WCSS)** for different values of kkk (number of clusters). The goal is to identify the point where adding more clusters doesn't significantly reduce WCSS.

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in C_j} ||x_j - \mu_i||^2$$

   where:

   (1) kkk = number of clusters
   (2) C_i = cluster i
   (3) X_j = data point in cluster i
   (4) μ_i = centroid of cluster i

   b) Silhouette Score: The Silhouette Score measures how well each data point fits within its assigned cluster. The score ranges from -1 to 1, where higher values indicate better-defined clusters.

$$s = \frac{b - a}{\max(a, b)}$$

   where,

   (1) a = average distance between the point and other points in the same cluster.

   (2) b = average distance between the point and points in the nearest cluster.

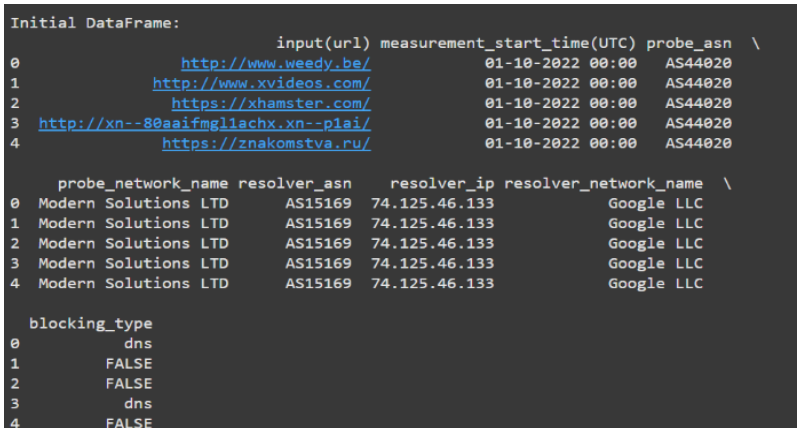
4) Based on the Elbow Method and Silhouette Scores, the optimal number of clusters k is selected.
5) K-Means Algorithm:
   a) Assign data points to k clusters.
   b) Minimize the distance between points and their cluster centroids.
   c) Iteratively updates the cluster centroids until convergence.
6) Dimensionality reduction using PCA for visualization.
7) Calculating Rand Index for validation.

4. Experimental Results

We primarily experimented with the datasets by combining the columns and keeping them separate to observe variations in accuracy. The dataset have been pre-processed separately for clustering and classification, Nomenclature for better understanding the datasets are:

- Original Dataset: This contains all 100,000 rows without formatting class labels in blocking_type column.
- Classification Dataset: This contains all 100,000 rows but with formatted class labels in bocking_type column (only 2 labels ‘true’, ‘false’)
- Clustering Dataset: This contains only the rows which contain non-false labels (only 3 labels ‘dns’, ‘http’, and ‘tcp-ip’)

4.1 Classification



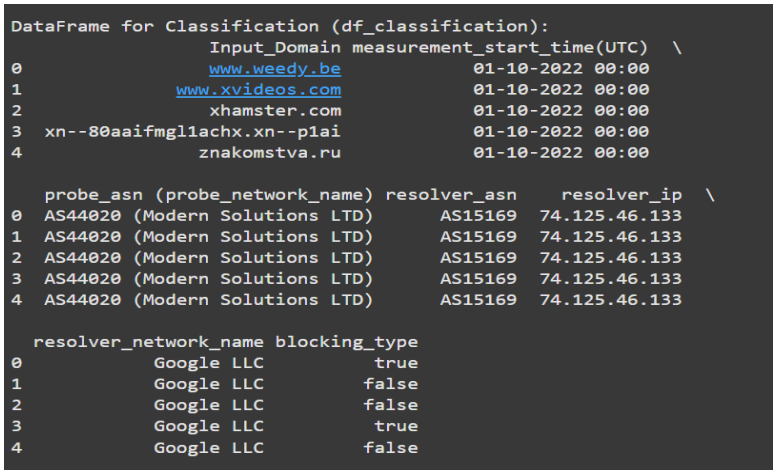
```
Initial DataFrame:
                          input(url) measurement_start_time(UTC) probe_asn  \
0                http://www.weedy.be/           01-10-2022 00:00   AS44020
1             http://www.xvideos.com/           01-10-2022 00:00   AS44020
2               https://xhamster.com/           01-10-2022 00:00   AS44020
3  http://xn--80aaifmgl1achx.xn--p1ai/           01-10-2022 00:00   AS44020
4               https://znakomstva.ru/           01-10-2022 00:00   AS44020

     probe_network_name resolver_asn    resolver_ip resolver_network_name  \
0  Modern Solutions LTD      AS15169  74.125.46.133           Google LLC
1  Modern Solutions LTD      AS15169  74.125.46.133           Google LLC
2  Modern Solutions LTD      AS15169  74.125.46.133           Google LLC
3  Modern Solutions LTD      AS15169  74.125.46.133           Google LLC
4  Modern Solutions LTD      AS15169  74.125.46.133           Google LLC

  blocking_type
0           dns
1         FALSE
2         FALSE
3           dns
4         FALSE
```

Figure 2: Original dataset without combining columns

1) Without combing any columns:



```
DataFrame for Classification (df_classification):
                Input_Domain measurement_start_time(UTC)  \
0                www.weedy.be           01-10-2022 00:00
1             www.xvideos.com           01-10-2022 00:00
2               xhamster.com            01-10-2022 00:00
3  xn--80aaifmgl1achx.xn--p1ai           01-10-2022 00:00
4               znakomstva.ru            01-10-2022 00:00

   probe_asn (probe_network_name) resolver_asn    resolver_ip  \
0  AS44020 (Modern Solutions LTD)      AS15169  74.125.46.133
1  AS44020 (Modern Solutions LTD)      AS15169  74.125.46.133
2  AS44020 (Modern Solutions LTD)      AS15169  74.125.46.133
3  AS44020 (Modern Solutions LTD)      AS15169  74.125.46.133
4  AS44020 (Modern Solutions LTD)      AS15169  74.125.46.133

  resolver_network_name blocking_type
0           Google LLC          true
1           Google LLC         false
2           Google LLC         false
3           Google LLC          true
4           Google LLC         false
```

Figure 3: Classification dataset without combining columns

Table 3: Fold Accuracies

Fold	Accuracy
1	0.969
2	0.959
3	0.949
4	0.962
5	0.948
Mean Accuracy	0.957

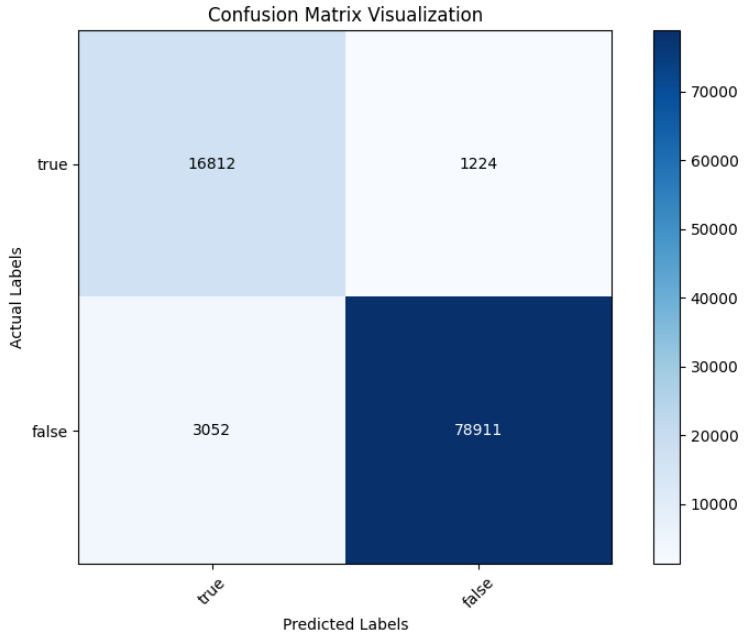


Figure 4: Confusion matrix for dataset without combining columns

2) Combining input(url), probe_asn and probe_network_name into one column Domain_Probe:

```
DataFrame for Classification (df_classification):
                                          Domain_Probe  \
0      www.weedy.be (AS44020 (Modern Solutions LTD))
1   www.xvideos.com (AS44020 (Modern Solutions LTD))
2      xhamster.com (AS44020 (Modern Solutions LTD))
3  xn--80aaifmgl1achx.xn--p1ai (AS44020 (Modern S...
4     znakomstva.ru (AS44020 (Modern Solutions LTD))

  measurement_start_time(UTC) resolver_asn    resolver_ip  \
0           01-10-2022 00:00      AS15169  74.125.46.133
1           01-10-2022 00:00      AS15169  74.125.46.133
2           01-10-2022 00:00      AS15169  74.125.46.133
3           01-10-2022 00:00      AS15169  74.125.46.133
4           01-10-2022 00:00      AS15169  74.125.46.133

  resolver_network_name blocking_type
0           Google LLC          true
1           Google LLC         false
2           Google LLC         false
3           Google LLC          true
4           Google LLC         false
```

Figure 5: Classification dataset with columns combined

Table 4: Fold Accuracies

Fold	Accuracy
1	0.937
2	0.929
3	0.933
4	0.900
5	0.915
Mean Accuracy	0.923

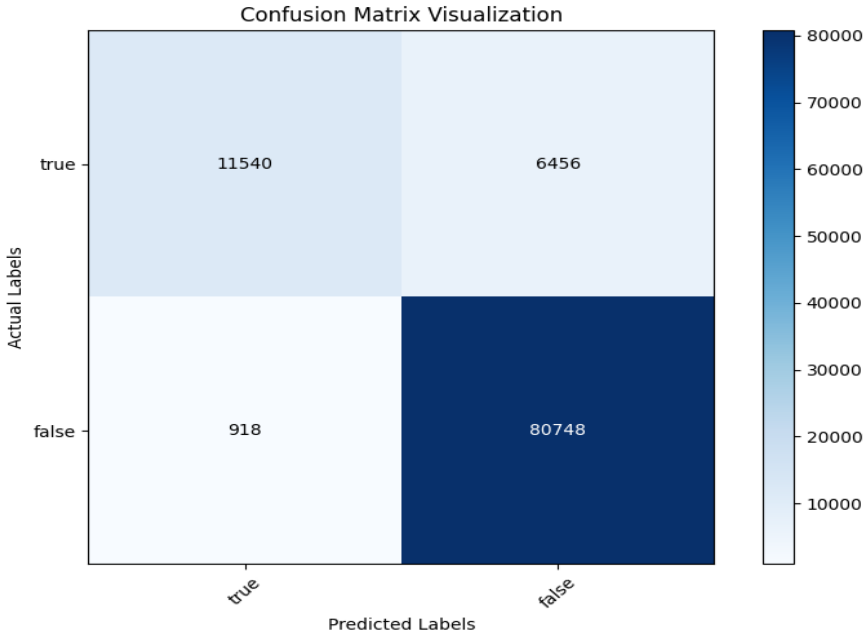


Figure 6: Confusion matrix for dataset with combined columns

4.2 Clustering

1) Without combining any columns:

	Input_Domain	day	quarter	probe_asn (probe_network_name)	\
0	www.weedy.be	1	Q1	AS44020 (Modern Solutions LTD)	
1	www.xvideos.com	1	Q1	AS44020 (Modern Solutions LTD)	
2	xhamster.com	1	Q1	AS44020 (Modern Solutions LTD)	
3	xn--80aaifmgl1achx.xn--p1ai	1	Q1	AS44020 (Modern Solutions LTD)	
4	znakomstva.ru	1	Q1	AS44020 (Modern Solutions LTD)	

	resolver_asn	resolver_ip	resolver_network_name	blocking_type
0	AS15169	74.125.46.133	Google LLC	dns
1	AS15169	74.125.46.133	Google LLC	false
2	AS15169	74.125.46.133	Google LLC	false
3	AS15169	74.125.46.133	Google LLC	dns
4	AS15169	74.125.46.133	Google LLC	false

Figure 7:Original dataset without combining columns with formatted time

a) With whole dataset:

Preview of the dataset:

	Input_Domain	day	quarter	probe_asn (probe_network_name)	\
0	-0.001969	-0.000382	0.002992	0.001589	
1	0.004265	-0.000382	0.002992	0.001589	
2	0.000081	-0.000382	0.002992	0.001589	
3	0.003432	-0.000382	0.002992	0.001589	
4	0.004067	-0.000382	0.002992	0.001589	

	resolver_asn	resolver_ip	resolver_network_name	blocking_type
0	0.000158	-0.000137	0.002542	dns
1	0.000158	-0.000137	0.002542	false
2	0.000158	-0.000137	0.002542	false
3	0.000158	-0.000137	0.002542	dns
4	0.000158	-0.000137	0.002542	false

Figure 8: Encoded original dataset without combining columns and with formatting time

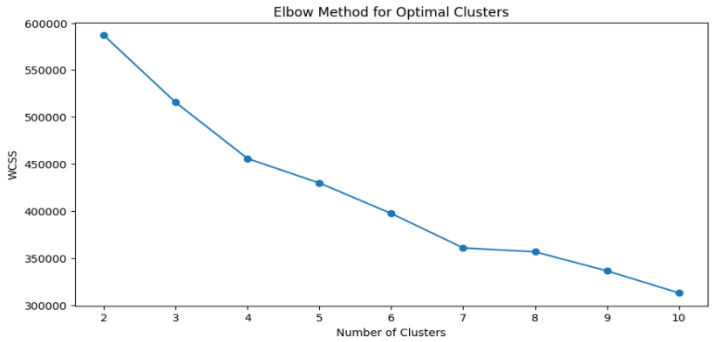


Figure 9: Elbow method for measuring optimal clusters

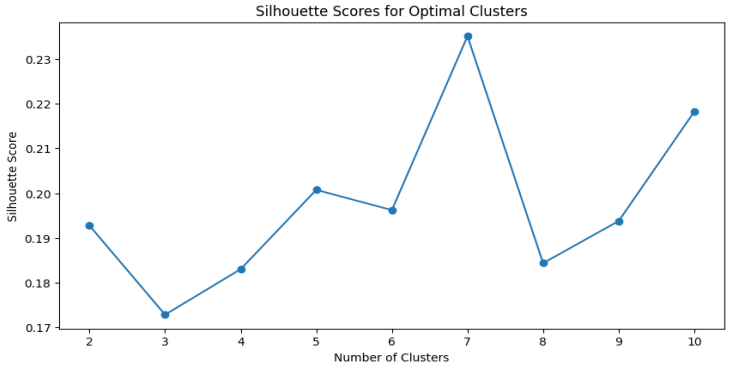


Figure 10: Silhouette scores for measuring optimal clusters

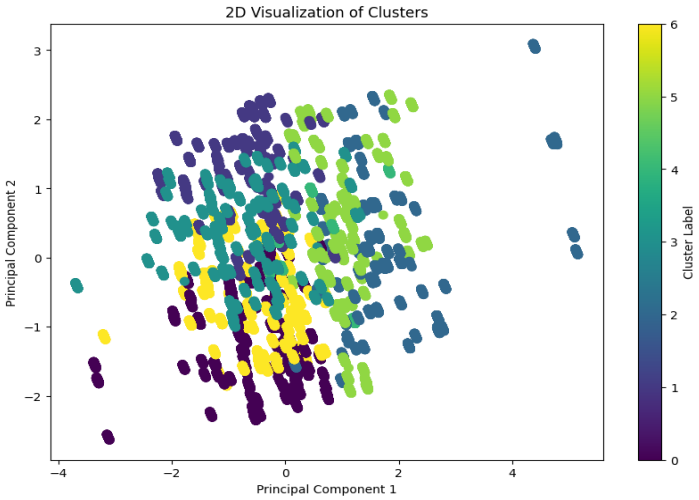


Figure 11: Clustering visualization

b) With only true labels in the dataset:
   i) With formatting measurement_start_time(UTC):

```
Preview of the dataset:
   Input_Domain      day   quarter  probe_asn (probe_network_name)  \
0      0.007006 -0.00086  0.009107                        0.006003
1     -0.001384 -0.00086  0.009107                        0.006003
2      0.000999 -0.00086  0.009107                       -0.006125
3      0.003042 -0.00086  0.009107                       -0.001622
4     -0.008149 -0.00086  0.009107                       -0.001622

   resolver_asn  resolver_ip  resolver_network_name blocking_type
0     -0.000099    -0.001831              -0.003403          dns
1     -0.000099    -0.001831              -0.003403          dns
2     -0.001132     0.011230               0.002185         http
3     -0.008350     0.005878              -0.000719          dns
4     -0.008350     0.005878              -0.000719          dns
```

Figure 12: Encoded clustering dataset without combining columns and with time formatting

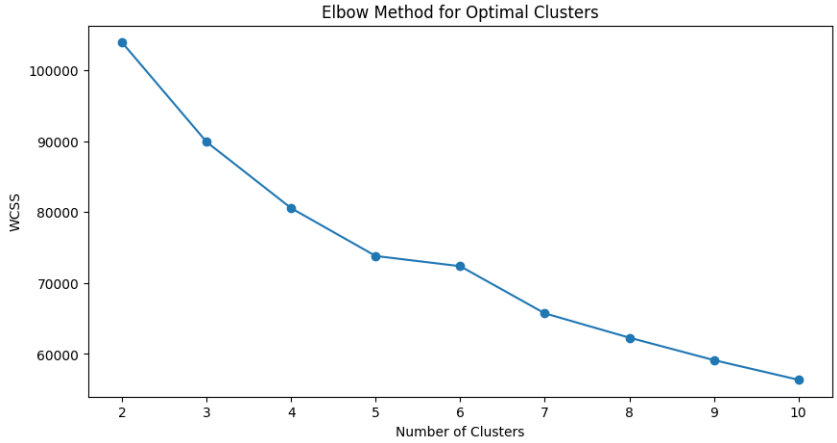


Figure 9: Elbow method for measuring optimal clusters

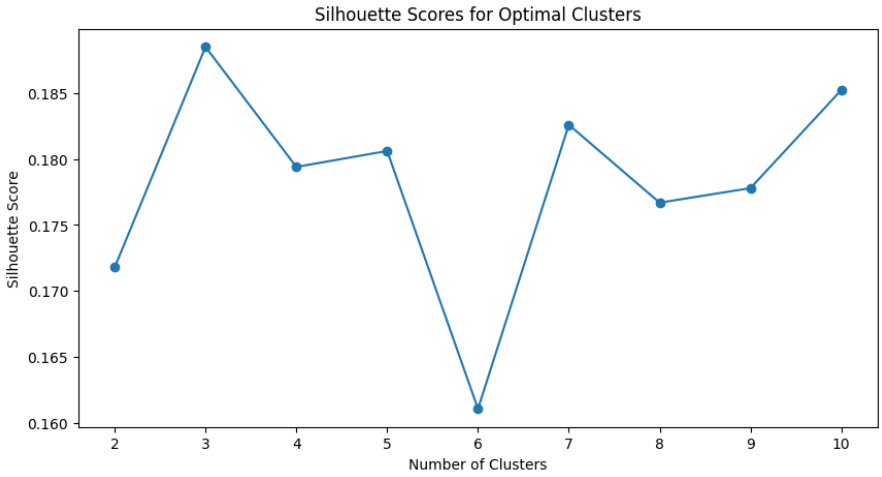


Figure 14: Silhoutte scores for measuring optimal clustering

Figure 10: Clustering visualization

2) Combining input(url), probe_asn and probe_network_name into one column Domain_Probe:
   a) With whole dataset:

	Domain_Probe	day	quarter	\
0	www.weedy.be (AS44020 (Modern Solutions LTD))	1	Q1	
1	www.xvideos.com (AS44020 (Modern Solutions LTD))	1	Q1	
2	xhamster.com (AS44020 (Modern Solutions LTD))	1	Q1	
3	xn--80aaifmgl1achx.xn--p1ai (AS44020 (Modern S...	1	Q1	
4	znakomstva.ru (AS44020 (Modern Solutions LTD))	1	Q1	

	resolver_asn	resolver_ip	resolver_network_name	blocking_type
0	AS15169	74.125.46.133	Google LLC	dns
1	AS15169	74.125.46.133	Google LLC	false
2	AS15169	74.125.46.133	Google LLC	false
3	AS15169	74.125.46.133	Google LLC	dns
4	AS15169	74.125.46.133	Google LLC	false

Figure 11: Original Dataset with combined columns

	Domain_Probe	day	quarter	\
0	www.weedy.be (AS44020 (Modern Solutions LTD))	1	Q1	
1	www.xvideos.com (AS44020 (Modern Solutions LTD))	1	Q1	
2	xhamster.com (AS44020 (Modern Solutions LTD))	1	Q1	
3	xn--80aaifmgl1achx.xn--p1ai (AS44020 (Modern S...	1	Q1	
4	znakomstva.ru (AS44020 (Modern Solutions LTD))	1	Q1	

	resolver_asn	resolver_ip	resolver_network_name	blocking_type
0	AS15169	74.125.46.133	Google LLC	dns
1	AS15169	74.125.46.133	Google LLC	false
2	AS15169	74.125.46.133	Google LLC	false
3	AS15169	74.125.46.133	Google LLC	dns
4	AS15169	74.125.46.133	Google LLC	false

Embedded dataset saved to 'global_embedded_dataset.csv'

Figure 17: Original dataset with time formatting

```
Preview of the dataset:
   Domain_Probe       day   quarter  resolver_asn  resolver_ip  \
0     -0.000066 -0.000382  0.003023     -0.000605     0.001779
1     -0.000118 -0.000382  0.003023     -0.000605     0.001779
2     -0.000112 -0.000382  0.003023     -0.000605     0.001779
3     -0.000043 -0.000382  0.003023     -0.000605     0.001779
4     -0.000109 -0.000382  0.003023     -0.000605     0.001779

   resolver_network_name blocking_type
0               0.004178           dns
1               0.004178         false
2               0.004178         false
3               0.004178           dns
4               0.004178         false
```

Figure 18: Encoded original dataset with time formatting

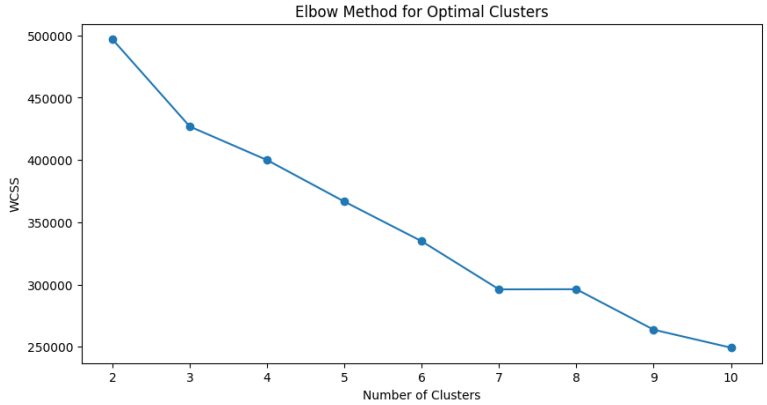


Figure 19: Elbow method for measuring optimal clusters

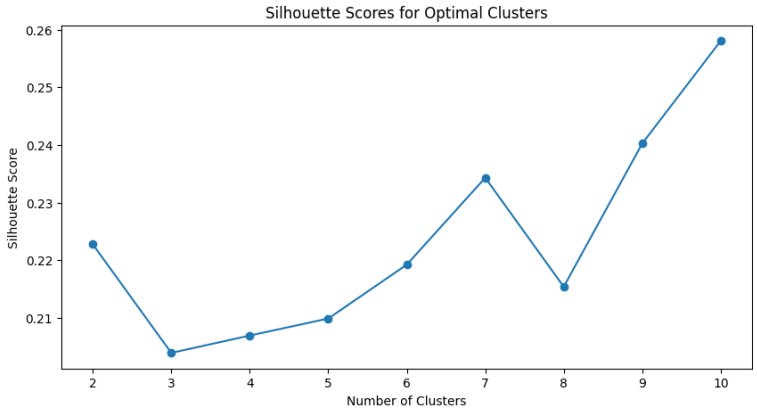


Figure 20: Silhoutte scores for measuring optimal clusters

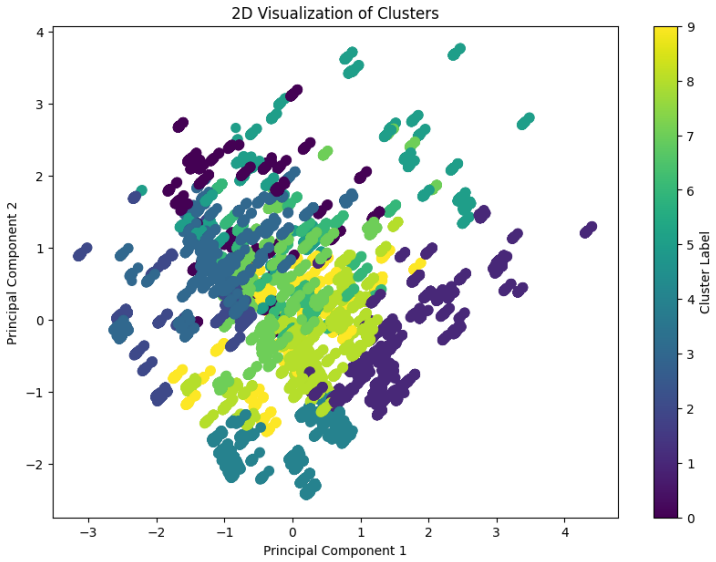


Figure 13: Clustering visualization

b) With only true labels in the dataset:

```
DataFrame for Clustering (df_clustering):
                                                Domain_Probe  \
0       www.weedy.be (AS44020 (Modern Solutions LTD))
3    xn--80aaifmgl1achx.xn--p1ai (AS44020 (Modern S...
10   www.facebook.com (AS15493 ("Russian company" L...
14           www.instagram.com (AS35807 (SkyNet Ltd.))
31                 twitter.com (AS35807 (SkyNet Ltd.))

   measurement_start_time(UTC) resolver_asn    resolver_ip  \
0             01-10-2022 00:00      AS15169  74.125.46.133
3             01-10-2022 00:00      AS15169  74.125.46.133
10            01-10-2022 00:00      AS13335   172.68.12.36
14            01-10-2022 00:00      AS35807    94.19.255.3
31            01-10-2022 00:00      AS35807    94.19.255.3

   resolver_network_name blocking_type
0             Google LLC           dns
3             Google LLC           dns
10      Cloudflare, Inc.          http
14           SkyNet Ltd.           dns
31           SkyNet Ltd.           dns
```

Figure 12: Clustering dataset with columns combined

```
                                    Domain_Probe  day quarter  \
0       www.weedy.be (AS44020 (Modern Solutions LTD))    1      Q1
1  xn--80aaifmgl1achx.xn--p1ai (AS44020 (Modern S...    1      Q1
2   www.facebook.com (AS15493 ("Russian company" L...    1      Q1
3            www.instagram.com (AS35807 (SkyNet Ltd.))    1      Q1
4                  twitter.com (AS35807 (SkyNet Ltd.))    1      Q1

  resolver_asn    resolver_ip resolver_network_name blocking_type
0      AS15169  74.125.46.133            Google LLC           dns
1      AS15169  74.125.46.133            Google LLC           dns
2      AS13335   172.68.12.36      Cloudflare, Inc.          http
3      AS35807    94.19.255.3           SkyNet Ltd.           dns
4      AS35807    94.19.255.3           SkyNet Ltd.           dns
```

Figure 23: Clustering dataset with time formatting

```
Preview of the dataset:
   Domain_Probe       day   quarter  resolver_asn  resolver_ip  \
0     -0.005649 -0.00086  0.00824      0.007115    -0.005533
1     -0.005593 -0.00086  0.00824      0.007115    -0.005533
2     -0.008705 -0.00086  0.00824      0.010300    -0.003972
3      0.006571 -0.00086  0.00824      0.001008     0.003545
4      0.006122 -0.00086  0.00824      0.001008     0.003545

   resolver_network_name blocking_type
0              -0.008581          dns
1              -0.008581          dns
2              -0.002291         http
3               0.001533          dns
4               0.001533          dns
```

Figure 24: Encoded clustered dataset with time formatting

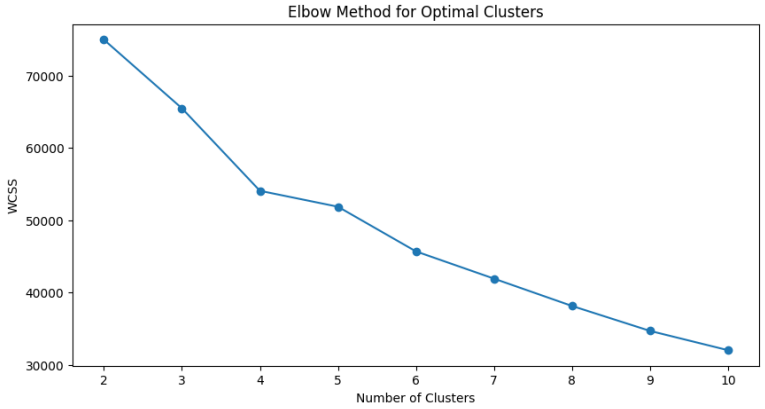


Figure 14: Elbow method for measuring optimal clusters

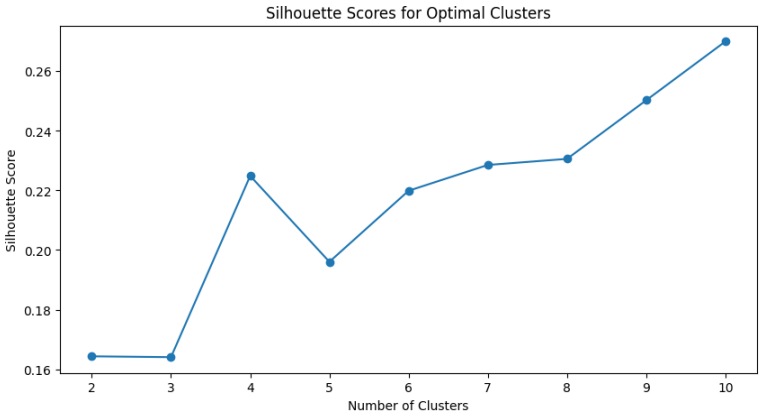


Figure 25: Silhouette scores for measuring optimal clusters

Table 5: Classification Metrices

	Mean Accuracy	**Time**
Combining Columns	92.3 %	40 minutes
Without Combining Columns	95.7 %	6 minutes

Table 6: Clustering Metrices

Column Status	Type of dataset	Time Formatting	Optimal Clusters	Rand Index	Time
Combined	Clustering	Yes	10	0.5577	1.25 minutes
Combined	Clustering	No	9	0.5507	1 minute
Combined	Original	Yes	10	0.3598	20 minutes
Combined	Original	No	9	0.3613	19 minutes
Not Combined	Clustering	Yes	3	0.5177	1.5 minutes
Not Combined	Clustering	No	9	0.5577	1.1 minutes
Not Combined	Original	Yes	7	0.3741	20 minutes
Not Combined	Original	No	2	0.5140	19 minutes

Finally, from the above 2 tables we could infer that, we were able to acquire good accuracy for classification. But for the clustering accuracy varies between low of medium.

5. Discussion

From this project, we learned about different censorship patterns and various types of censorship methods. we gained insights into data classification using decision trees and clustering with K-means clustering algorithm. Additionally, we understood the importance of preprocessing data sets, preparing data for model training and experimenting with different preprocessing techniques to enhance model accuracy.

If we were to redo this project, we would explore various classification techniques, evaluate their accuracies and compare the results to identify the most effective method. Similarly, we would experiment with different clustering algorithms and investigate others unsupervised learning techniques to improve clustering performance. In this project we got better classification accuracy, but the clustering accuracy was moderate so exploring different unsupervised learning techniques helps us to improve the clustering accuracy. We initially considered applying time series analysis to the data sets from October and November month to compare censorship patterns over time, however, due to data sets complexity we focused solely on October month data. In the future we would also incorporate correlation analysis to better understand the relationships between different features. This data set offers numerous opportunities for further exploration and deeper insights.

6. Conclusion

Regarding contributions to the project, Gnanitha primarily focused on programming for classification and clustering (60 %), while Jin concentrated on data preprocessing and writing tasks (40 %).

In conclusion, the project successfully achieved high accuracy in classifying censorship data, demonstrating the effectiveness of the decision tree model. However, clustering results showed varying accuracy, ranging from low to medium, indicating potential for improvement through alternative clustering methods. This analysis of censorship data mining provides valuable insights into censorship patterns and highlights the importance of exploring different techniques for better data interpretation.