BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

BRAC UNIVERSITY

Inspiring Excellence

Text Mining of News Articles to Detect Violation of Human Rights

AUTHORS
A M Saif Mahmud
Talha Ahmed
Azrin Hakim
Tanjida Sultana

SUPERVISOR
Dr. Md. Ashraful Alam
Assistant Professor
Department of CSE

**A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE**

**Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh**

December 2018

We would like to dedicate this thesis to our loving parents and respected faculty members . . .

Declaration

We hereby declare that this thesis is based on results obtained from our own work. All the materials that were used for the purpose of completing this thesis are duly acknowledged and mentioned in reference. This thesis, neither in whole nor in part, has been previously submitted to any other University or Institute for the award of any degree or diploma. We carried our research under the supervision of Dr. Md. Ashraful Alam.

Authors:

A M Saif Mahmud
Student ID: 15101045

Talha Ahmed
Student ID: 15101095

Azrin Hakim
Student ID: 15101049

Tanjida Sultana
Student ID: 15101013

Supervisor:

Dr. Md. Ashraful Alam
Assistant Professor, Department of Computer Science and Engineering
BRAC University

December 2018

The thesis titled Text Mining of News Articles to Detect Violation of Human Rights
Submitted by:
A M Saif Mahmud Student ID: 15101045
Talha Ahmed Student ID: 15101095
Azrin Hakim Student ID: 15101049
Tanjida Sultana Student ID: 15101013
of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Bachelor's Degree of Computer Science and Engineering.

1.	Md. Abdul Mottalib, PhD Professor and Chairperson BRAC University	Chairperson
2.	Dr. Md. Ashraful Alam Assistant Professor BRAC University	Supervisor
3.	Muhammad Abdur Rahman Adnan Lecturer BRAC University	Co-Supervisor

Acknowledgements

We would like to offer our sincere gratitude towards our thesis supervisor, Dr. Md. Ashraful Alam for his patient guidance and constant support for our work. In addition to that, we would also like to thank sir for his critical analysis and regular feedback of our work's progress, which guided us to remain in the aligned path to our goal. Furthermore, we would like to extend our gratitude to all the other faculty members of BRAC University Department of Computer Science and Engineering who played a vital part in the development process through the whole program.

We are also grateful to all the resources found on the internet which helped us to grasp a better understanding of machine learning. "Super Data Science", a website which contains machine learning tutorials, proved out to be really helpful along with the sites related to group forum and discussions.

Abstract

Violation of people's basic rights is a crime in almost every country in the world. However, many incidents in the real world gets passed unnoticed of legal attention. In this paper, we attempted to propose a model to detect violation of human rights, present in the incidents reported and stored in the form of news articles. Upon detecting the events of infringement, our model would classify a particular event to its specific type of breach, identify the location where the event has taken place and finally point out the geographical locations in a map for visualization. The system has been built on by using the news article archives of online news websites of The Daily Star, Prothom Alo and Dhaka Tribune.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

API	Application Programming Interface
DFS	Depth First Search
DOM	Document Object Model
EMM	Event-Based Media Monitoring
GADM	Database of Global Administrative Areas
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IDF	Inverse Document Frequency
NER	Named Entity Recognition
NGO	Non Governmental Organization
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NPHGS	Non-parametric Heterogenous Graph Scan
PHP	Hypertext Preprocessor
RDBMS	Relational Database Management System
SVM	Support Vector Machine
TF	Term Frequency

URL	Uniform Resource Locator
XPath	XML Path Language

Chapter 1

Introduction

Human Rights have been taken notice of ever since the first conceptualization of it by the natural law. However, it has not been a “really” pressing issue until the Universal Declaration of Human Rights took place back in 1948 [18]. Fortunately, Leaders of most of the nations endorse and elevate the laws passed by the declaration. Despite getting their support, nations still fail to provide proper and fair judgment.

In this age of technology, huge amount of information is available online. With the advent of online newspaper websites, keeping abreast with current situation in the world is made easy. For instance, Bangladesh has around 15 reliable online newspaper websites, enough to provide news coverage to almost every internet user. These news sources carry a lot of information about various topics, that are not being used.

In this paper, we endeavor to exploit the vast amount of information veiled in the form of news articles to quantitative data that is mobile and can be used for statistical analysis. By using the news archives of the top three online news websites in Bangladesh, namely, The Daily Star, Dhaka Tribune and Prothom Alo, we are going to detect the occurrences of human rights violation by extracting data from the news articles by using three different algorithms.

1.1 Motivation

Human Rights is defined as the fundamental rights a person is entitled to in every part of the world. In Bangladesh, the violation of human rights is higher than most countries and many Bengalis are deprived of human rights; mostly due to illiteracy, poverty and corruption. We seldom hear or read about the infringement of laws in newspapers or on TV. It has come to such a point in our lives that the violation of our rights as human has become something of a norm, and we as our country-men has excepted this as natural rather than questioning why we are not given such rights [7]. Hence, we are motivated in identifying the violation of our basic human rights via the tracing of different newspaper articles that are archived away, to acquire trends and patterns.

We have built a website that will give us an overview of the articles are in violation of human rights and indicating which human rights are being broken. The website will contain factual data and give the patterns in form of Heat map of the country.

By implementing data mining algorithms for information extraction from newspaper articles, we want our project to make use of this data to increase public awareness, help NGOs and Governments to take suitable actions against these infringements.

1.2 Objectives

Human rights in Bangladesh are cherished as central rights in Part III of the Constitution of Bangladesh. Be that as it may, protected and legitimate specialists trust a large number of the nation's laws expect change to uphold crucial rights and reflect majority rule estimations of the 21st century. This being said, people here are the most deprived of these rights.

To begin with, there are no records of human rights violation that keep occurring non-stop in every aspect of one's life. General people of the country are limited to very little knowledge of the basic rights one person is born with. To shed some light on such a significant issue, we will be using text mining algorithms to be able to categorize various different articles according to their respective type of violations in rights.

We will create a data bank where all the past data will be stored in order to have a proper track of the infringements that keep happening in almost all parts of Bangladesh. Typically, no trail is kept which gives people the chance to get away with all sorts of contravention. Our goal is to change this and create awareness in the society so that conventional people know more about the basic human rights.

1.3 Thesis Orientation

- **Chapter 2** contains Literature Review that provides the overall work flow, stating the algorithms and techniques used.
- **Chapter 3** discusses Proposed Model that gives a detailed description of every step of this model along with implementation details.
- **Chapter 4** shows the Experimental Setup and Results of all the algorithms used at each step of the model.
- **Chapter 5** concludes the paper while stating all the limitations faced and also discusses about future opportunities from this model.

Chapter 2

Literature Review

This section provides an overview of the related works done on the domain of Human Rights violation and text mining [23]. In addition to that, this section will also elaborate the background of text mining algorithms used in the proposed model.

As previously mentioned, ever since the United Nations have passed the Universal Declaration of Human Rights, the Governments and NGOs have started paying more attention towards the issue [4], to detect, analyze and predict the occurrences and causes of such events. Due to the advancement of technology, daily news is more dispersed and are readily available than ever before. In our research, we are going to make use of such news sources to contribute in the field of Human Rights, to detect, analyze, locate, educate and most importantly prevent such violations of basic human rights.

Researches have already done work in the area with various approaches. In recent times, organizations such as Amnesty International and Human Rights Watch are extensively monitoring the human rights event present in the data of social media [5]. Feng Chen and Daniel B. Neill of State University of New York at Albany have developed an algorithm called Non-parametric Heterogenous Graph Scan (NPHGS), which is used to detect human rights violation from event specific data from Twitter. The basic idea of the approach is to detect cluster of tweets of or pertaining to Human Rights by analyzing the key words. Then using a dictionary for Human Rights, tweets that were related were kept and the remaining were discarded.

In contrast to that, a team of researchers from the University of Minnesota have explored the use of event-based media monitoring in depth [1]. It describes two approaches of collecting data from the media, manually and automatically. It illustrates and differentiates manual and automatic EMM use which solely depends on the type of project or solution you are looking for.

In our project, we are going to use both manual data extraction for the training set and later on with the help of supervised learning, turn the whole system into automated process in other words, automatic extraction of data.

Chapter 3

Proposed Model

The proposed model is divided into five steps carrying out five different processes. Each process communicates through a common RDBMS provided by MariaDB. The five steps are:

- Article Scrapping
- Text Summarization
- Text Classification
- Location Extraction
- Visualization of Incidents

In this section, we are going to look into the steps of the proposed model in depth. Initially we are going to start by scrapping the articles off the internet from the publisher's websites. Then the system would summarize the articles to remove irrelevant sentences. The summarized articles are then passed on to text classifiers for categorizing the articles into their respective human rights point. After classification, the system extracts the location of the incidents from the articles. Once the above steps are done, data is represented in a heat map for visualization.

The data flowchart of the proposed model is presented below:

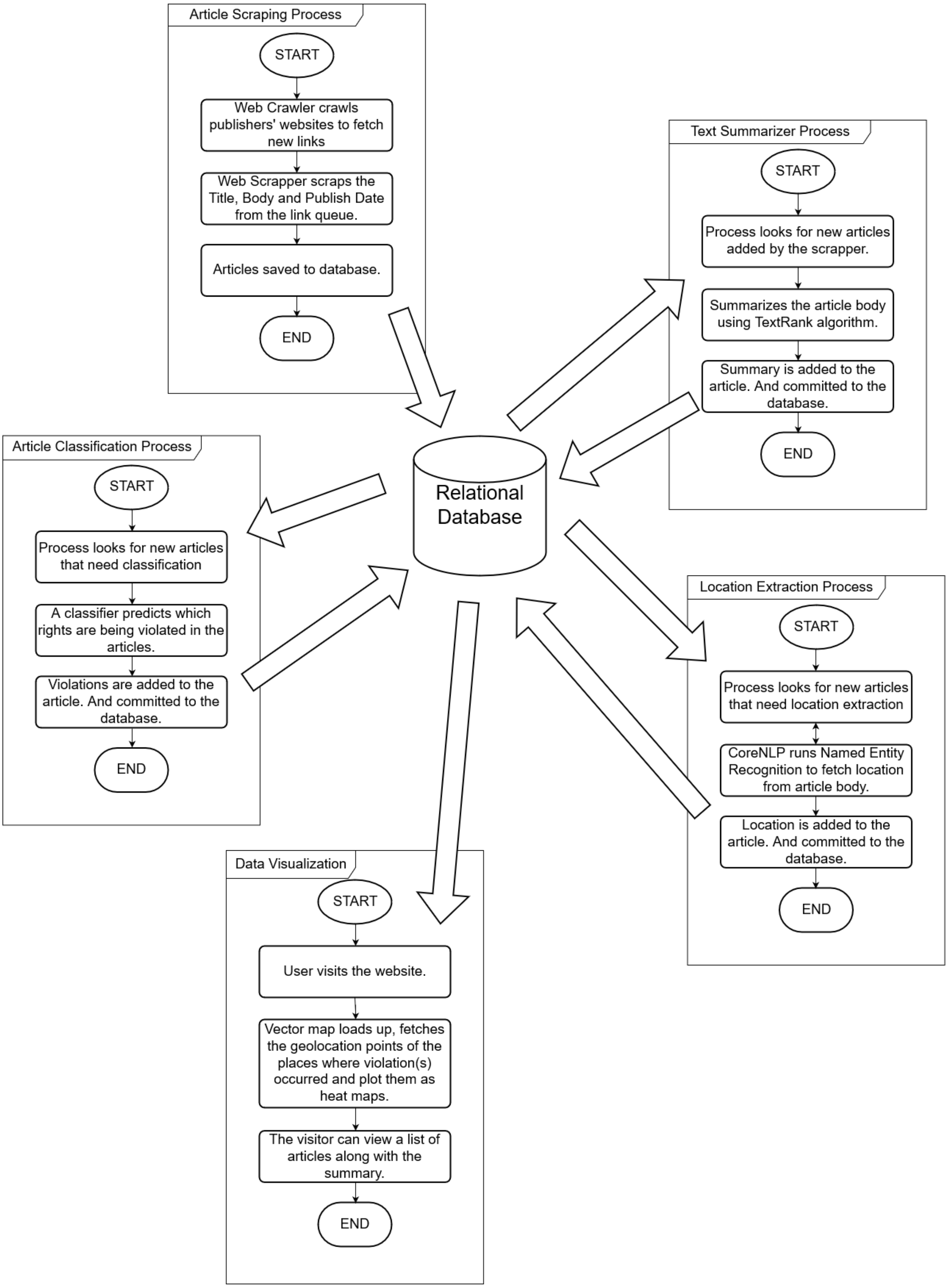

Article Scraping Process
START
Web Crawler crawls publishers' websites to fetch new links
Web Scrapper scraps the Title, Body and Publish Date from the link queue.
Articles saved to database.
END
Text Summarizer Process
START
Process looks for new articles added by the scrapper.
Summarizes the article body using TextRank algorithm.
Summary is added to the article. And committed to the database.
END
Article Classification Process
START
Process looks for new articles that need classification
A classifier predicts which rights are being violated in the articles.
Violations are added to the article. And committed to the database.
END
Relational Database
Location Extraction Process
START
Process looks for new articles that need location extraction
CoreNLP runs Named Entity Recognition to fetch location from article body.
Location is added to the article. And committed to the database.
END
Data Visualization
START
User visits the website.
Vector map loads up, fetches the geolocation points of the places where violation(s) occurred and plot them as heat maps.
The visitor can view a list of articles along with the summary.
END

Fig. 3.1 The workflow of the proposed model.

3.1 Web Crawling and Article Scrapping

To scrap the articles off of different news sources we are utilizing basic Web Crawling and Scraping techniques [3]. A Web Crawler [14] is a simple robot that fetches the Hyper Text Markup Language (HTML) code of a given base Uniform Resource Locator (URL) and collects the anchor links of the website's internal pages which are then added to a queue, and from here these links are processed by the crawler to fetch new sets of unique links from the HTML pages. We are using a Hypertext Preprocessor (PHP) library called DomCrawler by Symfony SAS for the project. This crawler utilizes Document Object Model Technology (DOM) to swiftly navigate through the web pages and collect the internal hyperlinks, and dump them in the database.

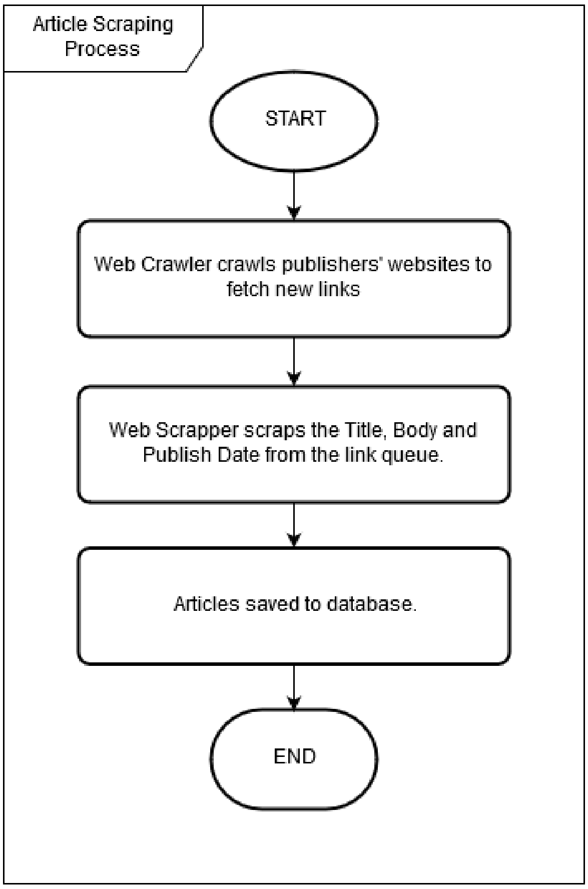

Article Scraping Process
START
Web Crawler crawls publishers' websites to fetch new links
Web Scrapper scraps the Title, Body and Publish Date from the link queue.
Articles saved to database.
END

Fig. 3.2 The web crawling and article scrapping process.

In combination with the Web Crawler, the Web Scrapper runs simultaneously, which re-fetches the HTML. Using XML Path Language (XPath), for each news portal that we scrap the pages from, we run specifically engineered queries to fetch the title, published date, and body. This information is then stored in the database for further processing.

3.2 Text Summarization

The articles that have been fetched by the scrapper are then passed into a separate process to summarize the articles into key sentences [11]. We are using the TextRank Algorithm [17] for this process, which is a graph-based ranking system that is used to calculate the importance of the words of each sentence. According to that information, the algorithm decides which sentences should be used for the summarization of the article.

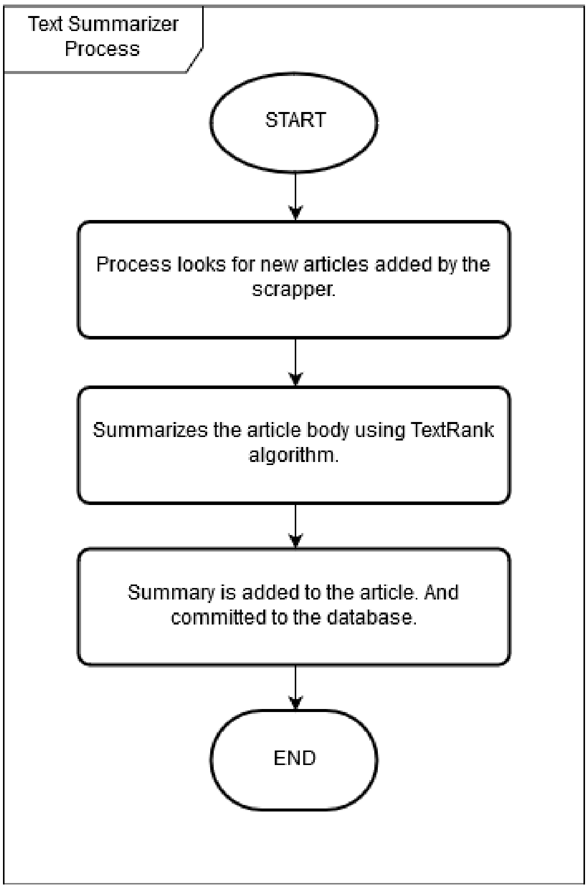

Text Summarizer Process
START
Process looks for new articles added by the scrapper.
Summarizes the article body using TextRank algorithm.
Summary is added to the article. And committed to the database.
END

Fig. 3.3 The text summarization process.

Each sentences are tokenized and parsed into a graph where the leaf nodes are the words. These leaf nodes are set with a default weight which change based on the number of occurrences of that particular word in the whole article. TFIDF [20] (short for **term frequency-inverse document frequency**) is a statistical function that measures the importance of a word in the document.

Term Frequency (or TF) measures the number of instances of the word present in a document.

$$tf(t,d) = 0.5 + 0.5 \bullet \frac{f_{t,d}}{\max\left\{f_{t,d} : t' \in d\right\}} \tag{3.1}$$

Inverse Document Frequency (or IDF) measures the importance of the word based on the repetition across multiple documents.

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{3.2}$$

The weight is then calculated based on the TF-IDF value that will increase depending on the number of times the word is present in the article. The more redundant the word is, the less weight it comprises of.

$$tfidf(t,d,D) = tf(t,d) \bullet idf(t,D) \tag{3.3}$$

Once the weight of each word is calculated, the sentences are then sorted based on the total weight the sentence accumulates. The top five sentences are selected from the sorted list as the summary of the article for our purpose.

A PHP library called TextRank, developed by PHPScience is used in the project for the purpose of summarizing the articles.

3.3 Text Classification

In the previous two stages, we have done the pre-processing of the articles: starting from scrapping the articles from the web pages using DOMCrawler to process the articles into a database. Furthermore, we summarized the data using TextRank algorithm, to get the main idea of an article – eliminating the irrelevant information. This is used to minimize redundant data to get a better understanding of the text for analysis.

In this stage we are going to use multiple algorithms to classify the news articles. A corpus [6] is built for each Human Right, which are used to determine whether an article falls in any of the categories, that can be attained by algorithms such as Naïve Bayes and Decision Tree, etc.

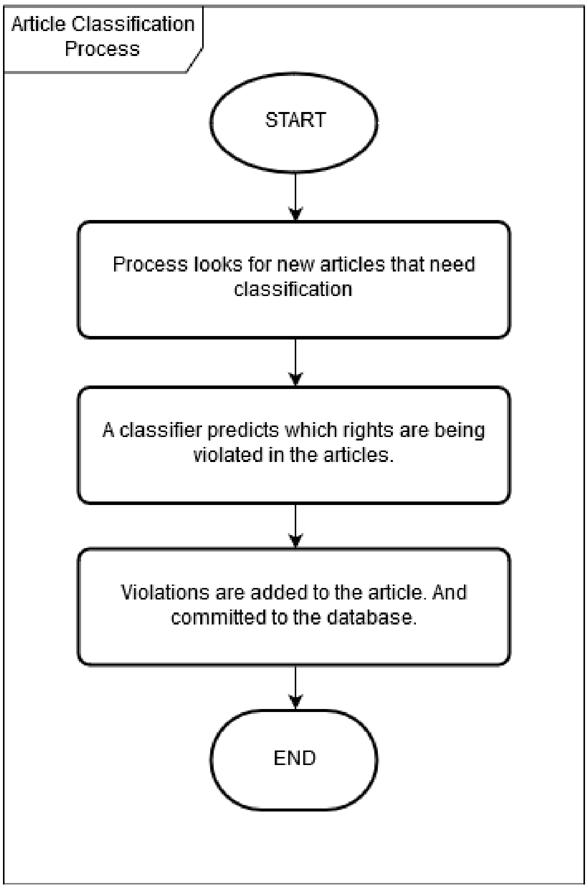

Article Classification Process
START
Process looks for new articles that need classification
A classifier predicts which rights are being violated in the articles.
Violations are added to the article. And committed to the database.
END

Fig. 3.4 The text classification process.

According to the Universal Declaration of Human Rights, there are thirty basic rights a human being is entitled to. We will check whether the thirty rights have been violated in the dataset. Upon detection, we will classify the articles to their respective rights.

This categorization is done by using seven different algorithms. The comparison of the algorithms will be determined, and the best result of a particular instance will be selected.

3.3.1 Naïve Bayes

Naïve Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naïve" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x(1) through x(n):

$$P(c \mid x) = \frac{P(x|c) \bullet P(c)}{P(x)} \tag{3.4}$$

We are going to test out two different types of Naïve Bayes classifiers:

- **Gaussian**: which works with normal distribution taking continuous features [9].

- **Bernoulli**: which assumes all features are in binary form [2].

3.3.2 Support Vector Machine

Support Vector Machine [12] is another robust classifier which relies on by a separating hyperplane. To put it simply, when training datasets are given, it produces an optimal hyperplane by using the labeled data to recognize new categories. SVM is also another example of supervised learning.

The figures below show an example of a hyperplane drawn for a simple dataset using SVM [19].

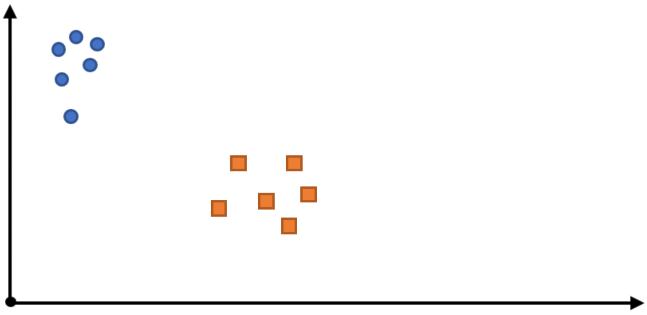


Fig. 3.5 Draw a line that separates blue circles and orange squares.

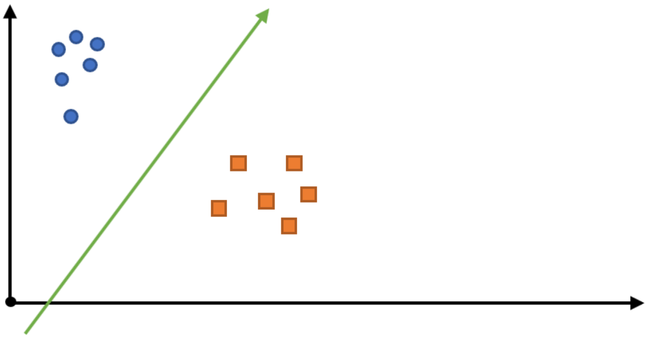


Fig. 3.6 Sample cut to divide into two classes.

3.3.3 Random Forest

It is another example of supervised learning. In Random Forest [22], as the name suggests, it creates a forest of decision trees, most of the time trained with ‘bagging” method. In other words, Random Forest merges multiple decision trees to get a more accurate and stable value. It selects the data randomly and select the best solution by means of voting.

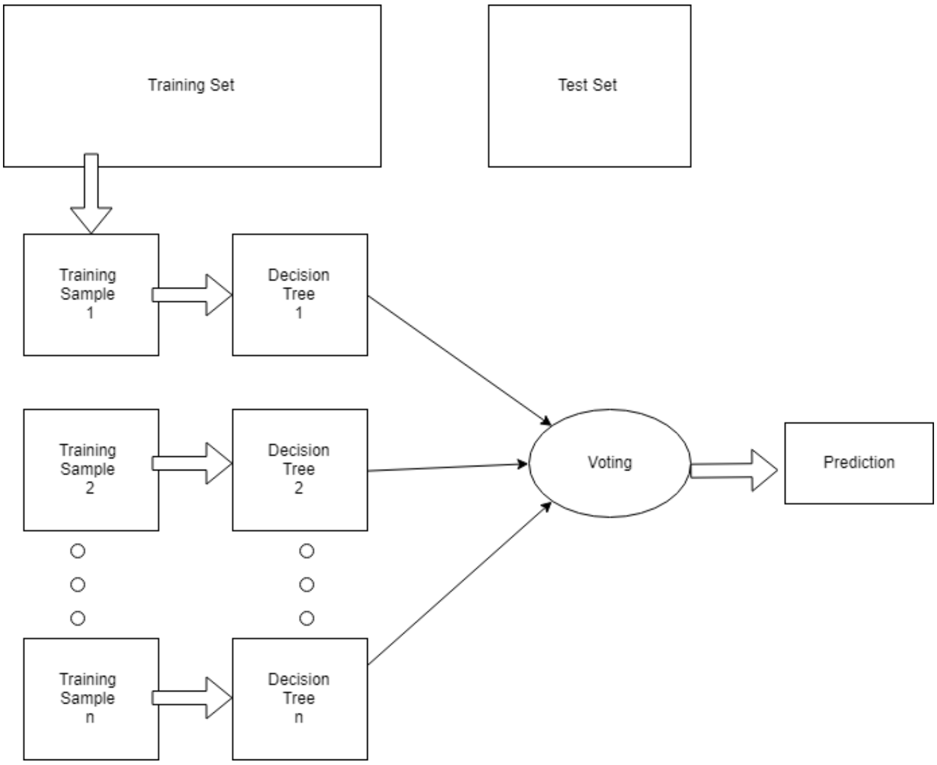


Fig. 3.7 The Random Forest process.

3.3.4 K-Nearest Neighbors

It is a simple algorithm which classifies new cases based on similarity measure [10]. A case is classified using majority vote of its neighbors measured by a distance function (where k is a positive integer, relatively small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

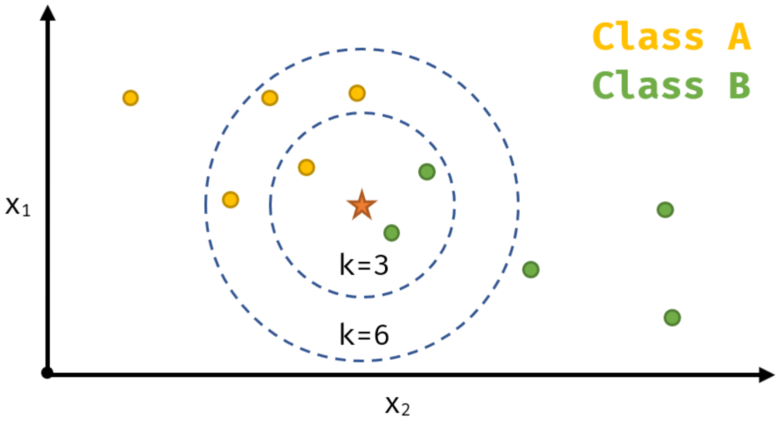


Fig. 3.8 Class of star would be B if k =3 and A if k = 6.

3.3.5 AdaBoost

AdaBoost [13] is short for Adaptive Boost, which retrains the algorithm recursively relying on the accuracy of previous training set. It is also a supervised classification algorithm. This equation briefly explains how AdaBoost work:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{3.5}$$

where, $h_t(x)$ is the output of weak classifier t for input x, and α_t is weight assigned to classifier.

3.3.6 Decision Tree

Decision Tree is a type of supervised learning algorithm which divides the data into a tree containing nodes and leaves depending on predefined parameters [24].

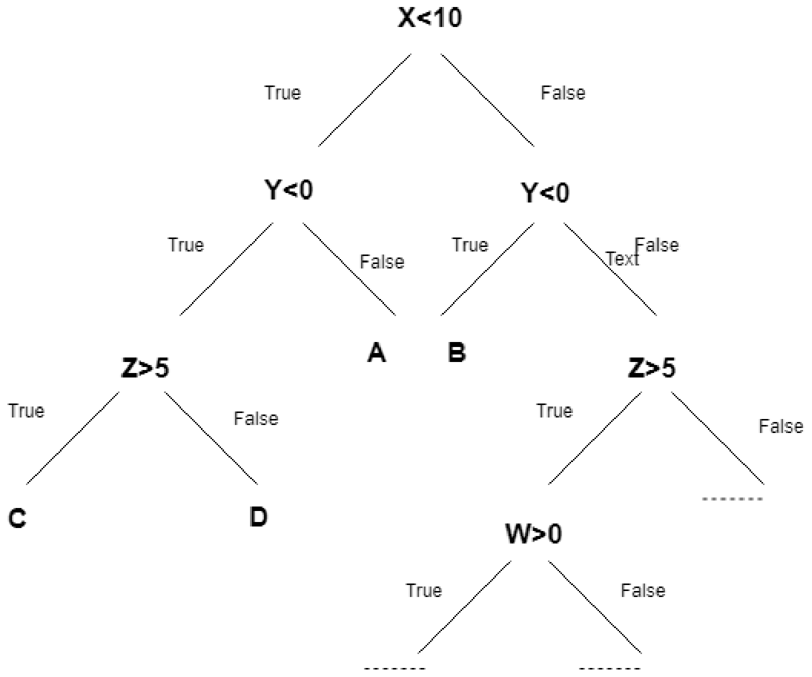


Fig. 3.9 How a tree is built according to the parameters.

3.4 Location Extraction

One of the vital aspects to create a record for analysis of any incident is to take account of the location where the event has taken place. In the project, we are going to extract the geographical location of the events that are breaching basic human rights. This is done by using the most enriched library of Stanford's CoreNLP [15], which has a robust Named Entity Recognition algorithm used for extracting locations. This is a good choice since it is very powerful compared to NLTK, OpenNLP and Pytorch.

The figure below shows the flowchart of the location extraction used in our model:

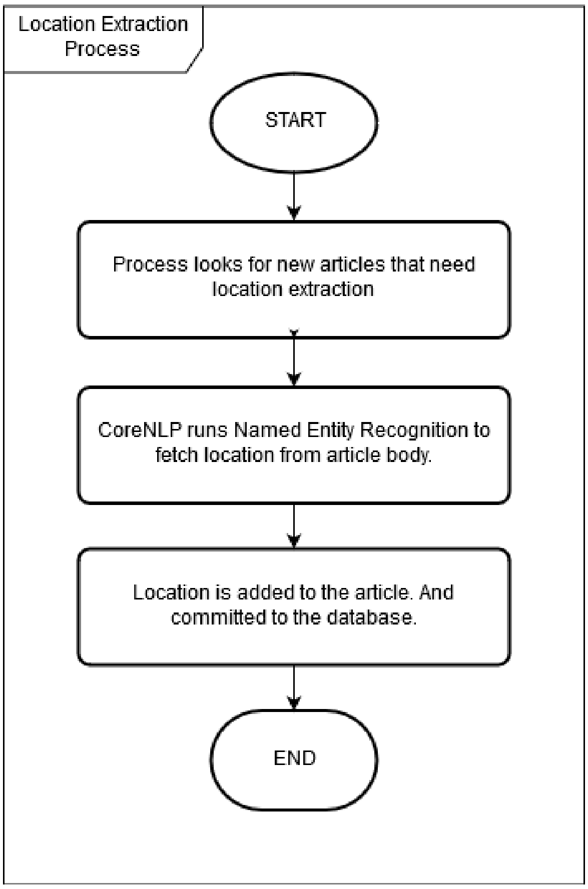

Location Extraction Process
START
Process looks for new articles that need location extraction
CoreNLP runs Named Entity Recognition to fetch location from article body.
Location is added to the article. And committed to the database.
END

Fig. 3.10 The location extraction process.

The extracted location of a specific article would then be stored into the database which is then going to be used for displaying in the heat map using mapbox.

3.5 Visualization of Incidents

After successfully classifying and extracting the location of our dataset, we are going to move on to displaying the results on the website.

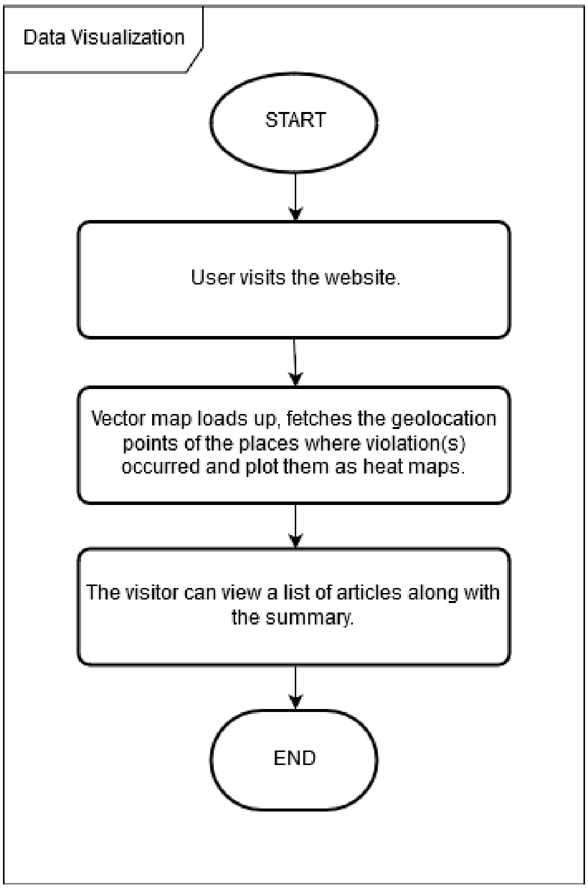

Data Visualization
START
User visits the website.
Vector map loads up, fetches the geolocation points of the places where violation(s) occurred and plot them as heat maps.
The visitor can view a list of articles along with the summary.
END

Fig. 3.11 The data visualization process.

The information that has been extracted so far from the articles are then rendered into a graphical form to give the users a visual stimulation for a better interpretation of the data. Since this project is a web based platform, we are using the WebGL technology to render Vector Maps onto a HTML5 canvas.

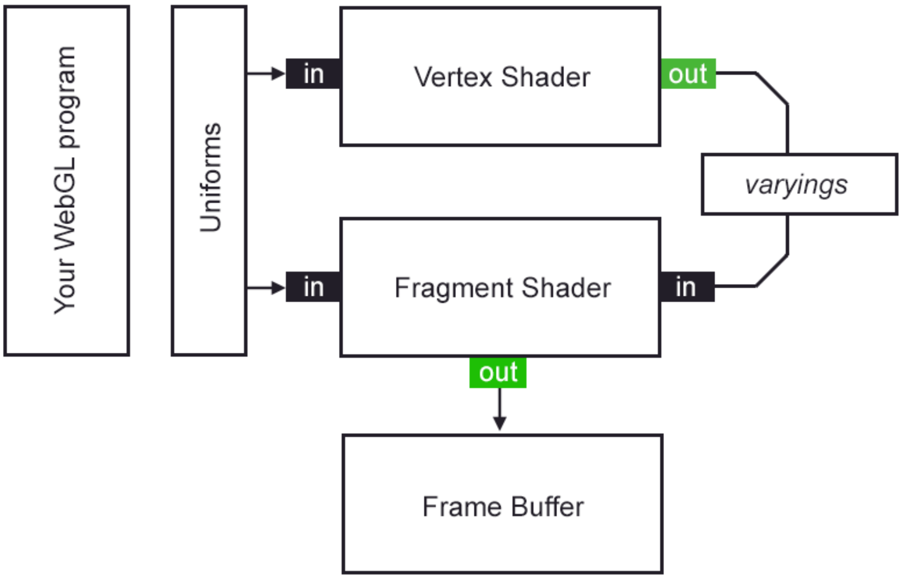


Fig. 3.12 The WebGL rendering pipeline. [21]

WebGL [16] is a variant of OpenGL supported by web browsers, which makes rendering computer graphics using JavaScript possible. OpenGL is an API which adds the ability for different programming languages to directly execute GPU commands as per the specifications.

A JavaScript module developed by mapbox [8], provides us with necessary APIs to build an interactive vector map using WebGL. The corresponding coordinates of the geo-location are fetched from the processed articles. A heatmap layer is added on top of the map to visualize the violation frequency by location.

Visitors would also be able to view the articles, including the rights violated and the summary, regarding the incidents occurred in the particular area. To view the list, the visitor simply has to click on the map and the system would generate a list of incidents within 15 kilometers radius of the point clicked.

Chapter 4

Experimental Setup and Result Analysis

4.1 Experimental Setup

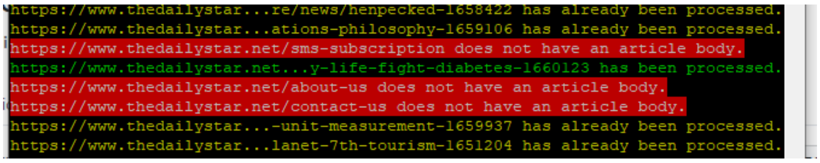
As mentioned above, our proposed model is divided into five steps, namely:

- Article Scraping
- Text Summarization
- Text Classification
- Location Extraction
- Visualization of Incidents

Each and every step follows the sequence of order to move on to the next step as shown in the flowchart and they are independent of each other. The dataset for each step is specific.

4.1.1 Web Crawling and Article Scrapping

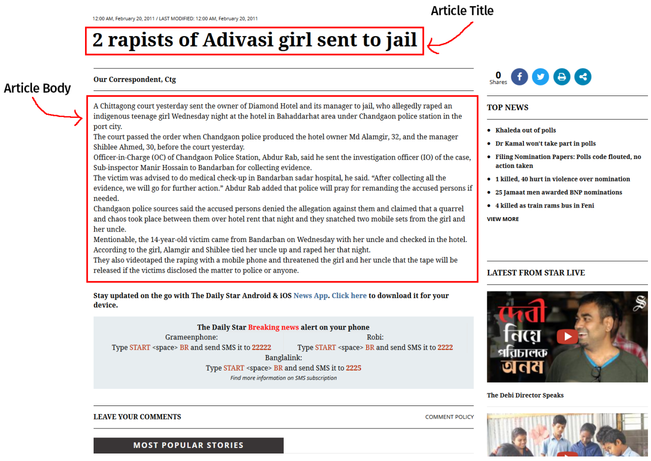
The articles have been taken from **The Daily Star**, **Prothom Alo** and **Dhaka Tribune** for scrapping and have been stored in the database according to the respective publisher, using a web crawler and scrapper.



```
https://www.thedailystar...re/news/henpecked-1658422 has already been processed.
https://www.thedailystar...ations-philosophy-1659106 has already been processed.
https://www.thedailystar.net/sms-subscription does not have an article body.
https://www.thedailystar.net...y-life-fight-diabetes-1660123 has been processed.
https://www.thedailystar.net/about-us does not have an article body.
https://www.thedailystar.net/contact-us does not have an article body.
https://www.thedailystar...-unit-measurement-1659937 has already been processed.
https://www.thedailystar...lanet-7th-tourism-1651204 has already been processed.
```

Fig. 4.1 Web crawler and scrapper in action.

As an example: for **The Daily Star**, we are fetching these by the scrapper and storing in the database for processing:



Article Title

Article Body

12:00 AM, February 20, 2011 / LAST MODIFIED: 12:00 AM, February 20, 2011

2 rapists of Adivasi girl sent to jail

Our Correspondent, Ctg

A Chittagong court yesterday sent the owner of Diamond Hotel and its manager to jail, who allegedly raped an indigenous teenage girl Wednesday night at the hotel in Bahaddarhat area under Chandgaon police station in the port city.

The court passed the order when Chandgaon police produced the hotel owner Md Alamgir, 32, and the manager Shiblee Ahmed, 30, before the court yesterday.

Officer-in-Charge (OC) of Chandgaon Police Station, Abdur Rab, said he sent the investigation officer (IO) of the case, Sub-inspector Manir Hossain to Bandarban for collecting evidence.

The victim was advised to do medical check-up in Bandarban sadar hospital, he said. “After collecting all the evidence, we will go for further action.” Abdur Rab added that police will pray for remanding the accused persons if needed.

Chandgaon police sources said the accused persons denied the allegation against them and claimed that a quarrel and chaos took place between them over hotel rent that night and they snatched two mobile sets from the girl and her uncle.

Mentionable, the 14-year-old victim came from Bandarban on Wednesday with her uncle and checked in the hotel. According to the girl, Alamgir and Shiblee tied her uncle up and raped her that night.

They also videotaped the raping with a mobile phone and threatened the girl and her uncle that the tape will be released if the victims disclosed the matter to police or anyone.

Stay updated on the go with The Daily Star Android & iOS News App. Click here to download it for your device.

The Daily Star Breaking news alert on your phone

Grameenphone:
Type START <space> BR and send SMS it to 22222

Robi:
Type START <space> BR and send SMS it to 2222

Banglalink:
Type START <space> BR and send SMS it to 2225

Find more information on SMS subscription

LEAVE YOUR COMMENTS COMMENT POLICY

MOST POPULAR STORIES

TOP NEWS

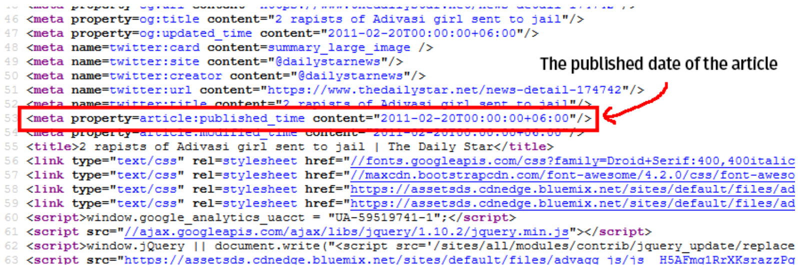
- Khaleda out of polls
- Dr Kamal won't take part in polls
- Filing Nomination Papers: Polls code flouted, no action taken
- 1 killed, 40 hurt in violence over nomination
- 25 Jamaat men awarded BNP nominations
- 4 killed as train rams bus in Feni

VIEW MORE

LATEST FROM STAR LIVE

The Debi Director Speaks

Fig. 4.2 Sections that contain the article title and body on The Daily Star.



```
<meta property=og:url content="https://www.thedailystar.net/news-detail-174742"/>
<meta property=og:title content="2 rapists of Adivasi girl sent to jail"/>
<meta property=og:updated_time content="2011-02-20T00:00:00+06:00"/>
<meta name=twitter:card content=summary_large_image />
<meta name=twitter:site content="@dailystarnews"/>
<meta name=twitter:creator content="@dailystarnews"/>
<meta name=twitter:url content="https://www.thedailystar.net/news-detail-174742"/>
<meta name=twitter:title content="2 rapists of Adivasi girl sent to jail"/>
<meta property=article:published_time content="2011-02-20T00:00:00+06:00"/>
<meta property=article:modified_time content="2011-02-20T00:00:00+06:00"/>
<title>2 rapists of Adivasi girl sent to jail | The Daily Star</title>
<link type="text/css" rel=stylesheet href="//fonts.googleapis.com/css?family=Droid+Serif:400,400italic
<link type="text/css" rel=stylesheet href="//maxcdn.bootstrapcdn.com/font-awesome/4.2.0/css/font-aweso
<link type="text/css" rel=stylesheet href="https://assetsds.cdnedge.bluemix.net/sites/default/files/ad
<link type="text/css" rel=stylesheet href="https://assetsds.cdnedge.bluemix.net/sites/default/files/ad
<script>window.google_analytics_uacct = "UA-59519741-1";</script>
<script src="//ajax.googleapis.com/ajax/libs/jquery/1.10.2/jquery.min.js"></script>
<script>window.jQuery || document.write("<script src='/sites/all/modules/contrib/jquery_update/replace
<script src="https://assetsds.cdnedge.bluemix.net/sites/default/files/advagg_js/js__H5AFmq1RrXKsrazzPq
```

The published date of the article

Fig. 4.3 The meta tag containing the publish date of the article.

4.1.2 Text Summarization

The summarized version of the articles are used to work for the classification in order to reduce the space and time complexity of the system and increase the overall efficiency. This would eliminate the redundant words, keeping only the keywords and the useful information, allowing the classification algorithms to work faster and better.

4.1.3 Text Classification

One of the main goals of this research is to identify whether there is any incident violating the human rights in the articles and subsequently find out which category or point (of thirty human rights) it is breaching. This is done by using classification algorithms which are trained using classified articles according to the thirty human rights. The dataset is then used to produce bag of words or corpus for an individual point. Then we are going to fit into seven different classification algorithms to get the best result.

Initially, we are simply going to categorize the articles into two sets of data:

1. **Articles that violate human rights**: The algorithm would simply look for incidents found in the articles that are violating basic rights and the types they fall under.

2. **Articles that do not violate human rights**: Since we are taking articles of all categories from the website of the news agencies, we are making sure that we do not miss out on any of the incidents that were unnoticed and imperceptive to catch. However, there would be articles that do not have any information of incidents where a right has been violated. They are simply identified and then kept for future use.

After we have done that, using only the articles that has human rights violation incident inside it, we are going to look for the location tag, that is the location where the incident has taken place.

1. **Articles that have location**: Location is going to be extracted by using a dataset of location of Bangladesh which expands up to level three. For instance, Bangladesh → Dhaka Division → Dhaka District → Motijheel, where Bangladesh is a Level Zero data and Motijheel is a Level Three. These articles would be added into the list for visualization.

2. **Articles that do not have location**: Articles that do not have any information about the location in the body, are not going to be added into the list for visualization in WebGL. However, they are not going to be discarded completely since they are can be used for improvements for the classifier.

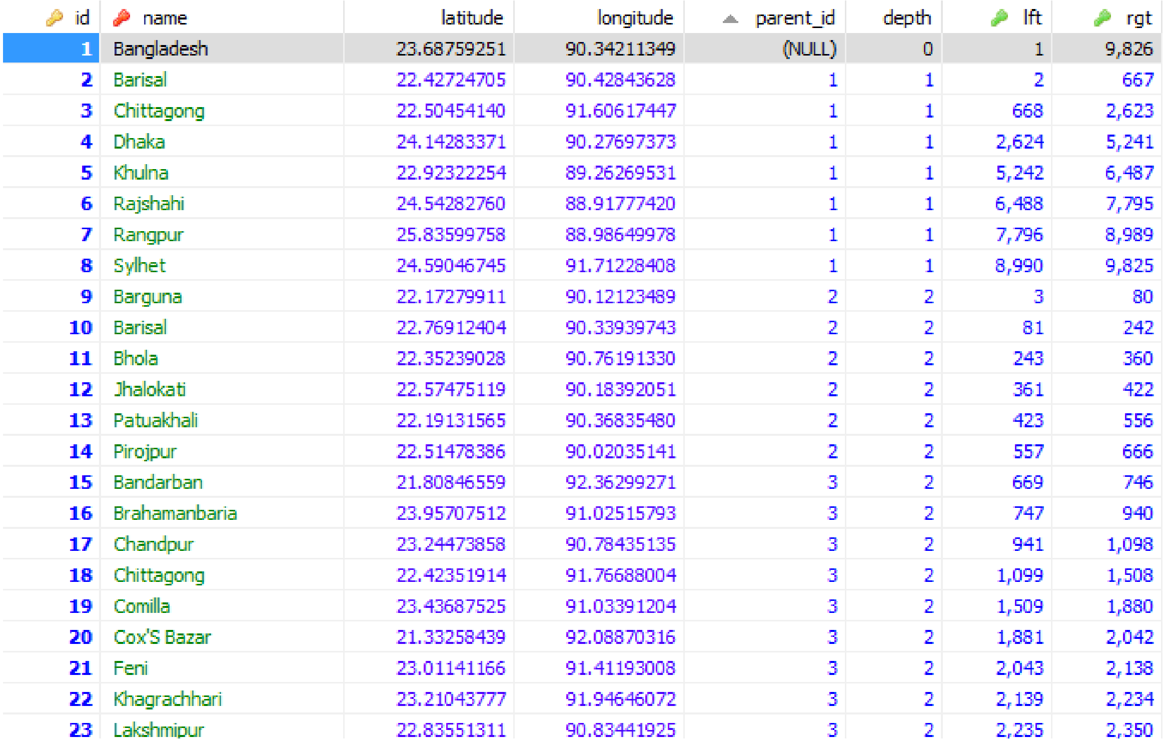
All the classification is done by using seven different algorithms:

- Naïve Bayes
  - Gaussian
  - Bernoulli
- Random Forest
- Support Vector Machine
- K-Nearest Neighbor
- AdaBoost
- Decision Tree

4.1.4 Location Extraction

As one of our objectives is to locate and display all the infringements in terms of geo-location, we have extracted the location from the articles using Stanford's CoreNLP. We have penetrated up to level three, as mentioned previously.

In order to plot geometric points on the vector map, we have seeded our database with GADM's level three geolocation information. As the data provided by GADM is flattened, the program had to run its own DFS algorithm to create a hierarchy tree of the locations.



id	name	latitude	longitude	parent_id	depth	lft	rgt
1	Bangladesh	23.68759251	90.34211349	(NULL)	0	1	9,826
2	Barisal	22.42724705	90.42843628	1	1	2	667
3	Chittagong	22.50454140	91.60617447	1	1	668	2,623
4	Dhaka	24.14283371	90.27697373	1	1	2,624	5,241
5	Khulna	22.92322254	89.26269531	1	1	5,242	6,487
6	Rajshahi	24.54282760	88.91777420	1	1	6,488	7,795
7	Rangpur	25.83599758	88.98649978	1	1	7,796	8,989
8	Sylhet	24.59046745	91.71228408	1	1	8,990	9,825
9	Barguna	22.17279911	90.12123489	2	2	3	80
10	Barisal	22.76912404	90.33939743	2	2	81	242
11	Bhola	22.35239028	90.76191330	2	2	243	360
12	Jhalokati	22.57475119	90.18392051	2	2	361	422
13	Patuakhali	22.19131565	90.36835480	2	2	423	556
14	Pirojpur	22.51478386	90.02035141	2	2	557	666
15	Bandarban	21.80846559	92.36299271	3	2	669	746
16	Brahamanbaria	23.95707512	91.02515793	3	2	747	940
17	Chandpur	23.24473858	90.78435135	3	2	941	1,098
18	Chittagong	22.42351914	91.76688004	3	2	1,099	1,508
19	Comilla	23.43687525	91.03391204	3	2	1,509	1,880
20	Cox'S Bazar	21.33258439	92.08870316	3	2	1,881	2,042
21	Feni	23.01141166	91.41193008	3	2	2,043	2,138
22	Khagrachhari	23.21043777	91.94646072	3	2	2,139	2,234
23	Lakshmipur	22.83551311	90.83441925	3	2	2,235	2,350

Fig. 4.4 Sample output data of the locations table.

The table have the following columns:

- **id** - An auto-generated unique identifier that will be used as the foreign key to link an article to the location.
- **name** - The name of the location.
- **latitude** and **longitude** - The geo-location coordinates of the location.
- **parent_id** - The identifier of the parent location, for example **Bangladesh** is the parent of **Dhaka Division** so **Dhaka Division** will have the identifier of **Bangladesh** in its **parent_id** column.

- **depth** - The level of the location (0 for country, 1 for region, 2 for administrative area, etc).
- **lft** and **rgt** - Auto-generated integers that is used to create a location hierarchy tree.

4.1.5 Visualization of Incidents

The dataset with the information of the incidents along with the specified locations are used to display in the heatmap for visuals.

4.2 Tools Used

The tools that were used in our project are listed below:

- JetBrains PhpStorm 2018 IDE
- JetBrains PyCharm 2018 IDE
- JetBrains IntelliJ IDEA 2018 IDE
- Git
- PHP 7.2
- Java 1.8
- Python 3.7
- Node.js 10
- MariaDB 10
- NPM
- Maven
- Composer
- XAMPP
- SASS

4.2.1 Libraries for PHP

- Laravel Framework
- GeoPHP
- TextRank by PHPScience
- GeoJSON
- DOMCrawler by Symfony SAS

4.2.2 Libraries for Java

- CoreNLP
- MySQL Connector

4.2.3 Libraries for Python

- NLTK
- NumPy
- Pandas
- sci-kit learn

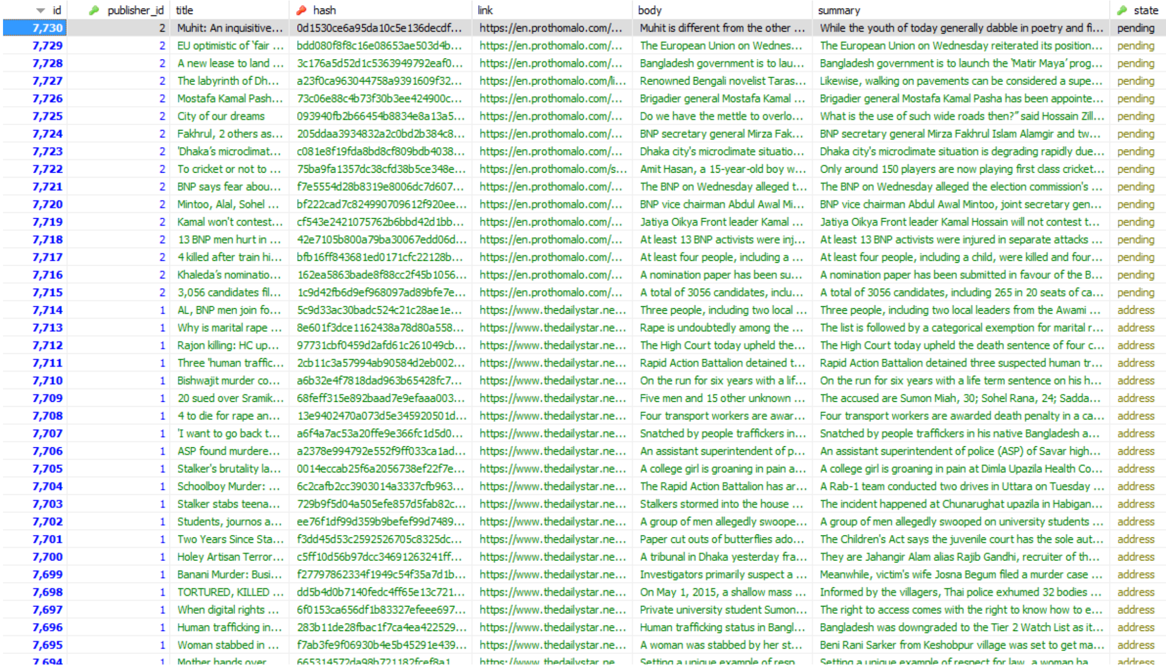
4.2.4 Libraries for JavaScript

- VueJS
- axios
- Semantic UI
- moment.js
- mapbox

4.3 Results

Once we have implemented our proposed model successfully, the whole process is simulated using the datasets that have been produced. As previously mentioned, we have five sequential stages, each stage has been tested separately to analyze and compare the results.

4.3.1 Dataset Used



id	publisher_id	title	hash	link	body	summary	state
7,730	2	Muhit: An inquisitive...	0d1530ce6a95da10c5e136decdf...	https://en.prothomalo.com/...	Muhit is different from the other ...	While the youth of today generally dabble in poetry and fi...	pending
7,729	2	EU optimistic of 'fair ...	bdd080f8f8c16e08653ae503d4b...	https://en.prothomalo.com/...	The European Union on Wednes...	The European Union on Wednesday reiterated its position...	pending
7,728	2	A new lease to land ...	3c176a5d52d1c5363949792eaf0...	https://en.prothomalo.com/...	Bangladesh government is to lau...	Bangladesh government is to launch the 'Matir Maya' prog...	pending
7,727	2	The labyrinth of Dh...	a23f0ca963044758a9391609f32...	https://en.prothomalo.com/li...	Renowned Bengali novelist Taras...	Likewise, walking on pavements can be considered a supe...	pending
7,726	2	Mostafa Kamal Pash...	73c06e88c4b73f30b3ee424900c...	https://en.prothomalo.com/...	Brigadier general Mostafa Kamal ...	Brigadier general Mostafa Kamal Pasha has been appointe...	pending
7,725	2	City of our dreams	093940fb2b66454b8834e8a13a5...	https://en.prothomalo.com/...	Do we have the mettle to overlo...	What is the use of such wide roads then?" said Hossain Zill...	pending
7,724	2	Fakhrul, 2 others as...	205ddaa3934832a2c0bd2b384c8...	https://en.prothomalo.com/...	BNP secretary general Mirza Fak...	BNP secretary general Mirza Fakhrul Islam Alamgir and tw...	pending
7,723	2	'Dhaka's microclimat...	c081e8f19fda8bd8cf809bdb4038...	https://en.prothomalo.com/...	Dhaka city's microclimate situatio...	Dhaka city's microclimate situation is degrading rapidly due...	pending
7,722	2	To cricket or not to ...	75ba9fa1357dc38cfd38b5ce348e...	https://en.prothomalo.com/s...	Amit Hasan, a 15-year-old boy w...	Only around 150 players are now playing first class cricket...	pending
7,721	2	BNP says fear abou...	f7e5554d28b8319e8006dc7d607...	https://en.prothomalo.com/...	The BNP on Wednesday alleged t...	The BNP on Wednesday alleged the election commission's ...	pending
7,720	2	Mintoo, Alal, Sohel ...	bf222cad7c824990709612f920ee...	https://en.prothomalo.com/...	BNP vice chairman Abdul Awal Mi...	BNP vice chairman Abdul Awal Mintoo, joint secretary gen...	pending
7,719	2	Kamal won't contest...	cf543e2421075762b6bbd42d1bb...	https://en.prothomalo.com/...	Jatiya Oikya Front leader Kamal ...	Jatiya Oikya Front leader Kamal Hossain will not contest t...	pending
7,718	2	13 BNP men hurt in ...	42e7105b800a79ba30067edd06d...	https://en.prothomalo.com/...	At least 13 BNP activists were inj...	At least 13 BNP activists were injured in separate attacks ...	pending
7,717	2	4 killed after train hi...	bfb16ff843681ed0171cfc22128b...	https://en.prothomalo.com/...	At least four people, including a ...	At least four people, including a child, were killed and four...	pending
7,716	2	Khaleda's nominatio...	162ea5863bade8f88cc2f45b1056...	https://en.prothomalo.com/...	A nomination paper has been su...	A nomination paper has been submitted in favour of the B...	pending
7,715	2	3,056 candidates fil...	1c9d42fb6d9ef968097ad89bfe7e...	https://en.prothomalo.com/...	A total of 3056 candidates, inclu...	A total of 3056 candidates, including 265 in 20 seats of ca...	pending
7,714	1	AL, BNP men join fo...	5c9d33ac30badc524c21c28ae1e...	https://www.thedailystar.ne...	Three people, including two local ...	Three people, including two local leaders from the Awami ...	address
7,713	1	Why is marital rape ...	8e601f3dce1162438a78d80a558...	https://www.thedailystar.ne...	Rape is undoubtedly among the ...	The list is followed by a categorical exemption for marital r...	address
7,712	1	Rajon killing: HC up...	97731cbf0459d2afd61c261049cb...	https://www.thedailystar.ne...	The High Court today upheld the...	The High Court today upheld the death sentence of four c...	address
7,711	1	Three 'human traffic...	2cb11c3a57994ab90584d2eb002...	https://www.thedailystar.ne...	Rapid Action Battalion detained t...	Rapid Action Battalion detained three suspected human tr...	address
7,710	1	Bishwajit murder co...	a6b32e4f7818dad963b65428fc7...	https://www.thedailystar.ne...	On the run for six years with a lif...	On the run for six years with a life term sentence on his h...	address
7,709	1	20 sued over Sramik...	68feff315e892baad7e9efaaa003...	https://www.thedailystar.ne...	Five men and 15 other unknown ...	The accused are Sumon Miah, 30; Sohel Rana, 24; Sadda...	address
7,708	1	4 to die for rape an...	13e9402470a073d5e34592050 1d...	https://www.thedailystar.ne...	Four transport workers are awar...	Four transport workers are awarded death penalty in a ca...	address
7,707	1	'I want to go back t...	a6f4a7ac53a20ffe9e366fc1d5d0...	https://www.thedailystar.ne...	Snatched by people traffickers in...	Snatched by people traffickers in his native Bangladesh a...	address
7,706	1	ASP found murdere...	a2378e994792e552f9ff033ca1ad...	https://www.thedailystar.ne...	An assistant superintendent of p...	An assistant superintendent of police (ASP) of Savar high...	address
7,705	1	Stalker's brutality la...	0014eccab25f6a2056738ef22f7e...	https://www.thedailystar.ne...	A college girl is groaning in pain a...	A college girl is groaning in pain at Dimla Upazila Health Co...	address
7,704	1	Schoolboy Murder: ...	6c2cafb2cc3903014a3337cfb963...	https://www.thedailystar.ne...	The Rapid Action Battalion has ar...	A Rab-1 team conducted two drives in Uttara on Tuesday ...	address
7,703	1	Stalker stabs teena...	729b9f5d04a505efe857d5fab82c...	https://www.thedailystar.ne...	Stalkers stormed into the house ...	The incident happened at Chunarughat upazila in Habigan...	address
7,702	1	Students, journos a...	ee76f1df99d359b9befef99d7489...	https://www.thedailystar.ne...	A group of men allegedly swoope...	A group of men allegedly swooped on university students ...	address
7,701	1	Two Years Since Sta...	f3dd45d53c2592526705c8325dc...	https://www.thedailystar.ne...	Paper cut outs of butterflies ado...	The Children's Act says the juvenile court has the sole aut...	address
7,700	1	Holey Artisan Terror...	c5ff10d56b97dcc3469126324 1ff...	https://www.thedailystar.ne...	A tribunal in Dhaka yesterday fra...	They are Jahangir Alam alias Rajib Gandhi, recruiter of th...	address
7,699	1	Banani Murder: Busi...	f27797862334f1949c54f35a7d1b...	https://www.thedailystar.ne...	Investigators primarily suspect a ...	Meanwhile, victim's wife Josna Begum filed a murder case ...	address
7,698	1	TORTURED, KILLED ...	dd5b4d0b7140fedc4ff65e13c721...	https://www.thedailystar.ne...	On May 1, 2015, a shallow mass ...	Informed by the villagers, Thai police exhumed 32 bodies ...	address
7,697	1	When digital rights ...	6f0153ca656df1b83327efeee697...	https://www.thedailystar.ne...	Private university student Sumon...	The right to access comes with the right to know how to e...	address
7,696	1	Human trafficking in...	283b11de28fbac1f7ca4ea422529...	https://www.thedailystar.ne...	Human trafficking status in Bangl...	Bangladesh was downgraded to the Tier 2 Watch List as it...	address
7,695	1	Woman stabbed in ...	f7ab3fe9f06930b4e5b45291e439...	https://www.thedailystar.ne...	A woman was stabbed by her st...	Beni Rani Sarker from Keshobpur village was set to get ma...	address
7,694	1	Mother hands over	665314572da98b721182fcef8a1	https://www.thedailystar.ne	Setting a unique example of resp	Setting a unique example of respect for law, a woman ha	address

Fig. 4.5 Sample output of the scrapped articles.

The *articles* table contains the following columns:

- **id** - An auto-generated unique identifier of an article to build relations against.
- **publisher_id** - Contains the foreign key of the publisher's uniquely generated identifier.
- **title** - The title of the article.
- **hash** - A unique string created from the SHA-256 cryptographic hash algorithm against the original link of the article to make sure the article is unique.
- **link** - The link to the original source of the article.
- **body** - Contains the original article body to be processed by the **Text Summarizer** and **Location Extractor**.

- **summary** - Contains the summarized version of the article to be viewed by the users and the **Text Classifiers**.
- **state** - The current state of the article that determines which stage the article will go through next.
- **location_id** - Contains the foreign key of the location the incident took place.
- **published_at** - The date and time the article was originally published on the publisher's website.

4.3.2 Classification Results and Analysis

By testing the classifiers using the dataset, we have found the following results:

Table 4.1 Performance metrics of different classifiers.

Classifier	Accuracy	Log Loss (10^{-15})
Gaussian Naïve Bayes	80.952	6.579
Bernoulli Naïve Bayes	74.830	8.693
Random Forest	79.252	7.166
Support Vector Machine	86.054	4.817
K-Nearest Neighbor	74.830	8.694
AdaBoost	77.891	7.636
Decision Tree	81.020	6.555

As shown in the table above, we have calculated the accuracy and the **log-loss** of each classifier. Logarithmic Loss (or log-loss) is the measure of the performance of a classification model. It increases as the predicted probability differs from the actual value.

For each human right, we have fitted all the seven classifiers and found that each classifier gives the best result for a different violation type: which depends on the data size of the given class. Here are the best results:

Table 4.2 Classification report of Gaussian Naïve Bayes for "No Torture".

	Precision	Recall	F1-score	Support
0	0.81	0.97	0.89	76
1	0.71	0.23	0.34	22
Micro Average	0.81	0.81	0.81	98
Macro Average	0.76	0.60	0.62	98
Weighted Average	0.79	0.81	0.76	98

Table 4.3 Classification report of Bernoulli Naïve Bayes for "No Torture".

	Precision	Recall	F1-score	Support
0	0.78	1.00	0.87	76
1	0.00	0.00	0.00	22
Micro Average	0.78	0.78	0.78	98
Macro Average	0.39	0.50	0.44	98
Weighted Average	0.60	0.78	0.68	98

Table 4.4 Classification report of Random Forest for "The Right To Life".

	Precision	Recall	F1-score	Support
0	0.81	0.97	0.88	70
1	0.86	0.43	0.57	28
Micro Average	0.82	0.82	0.82	98
Macro Average	0.83	0.70	0.73	98
Weighted Average	0.82	0.82	0.79	98

Table 4.5 Classification report of SVM for "The Right To Life".

	Precision	Recall	F1-score	Support
0	0.90	0.94	0.92	70
1	0.84	0.75	0.79	28
Micro Average	0.89	0.89	0.89	98
Macro Average	0.87	0.85	0.86	98
Weighted Average	0.89	0.89	0.89	98

Table 4.6 Classification report of K-Nearest Neighbor for "No Torture".

	Precision	Recall	F1-score	Support
0	0.81	0.97	0.89	76
1	0.71	0.23	0.34	22
Micro Average	0.81	0.81	0.81	98
Macro Average	0.76	0.60	0.62	98
Weighted Average	0.79	0.81	0.76	98

Table 4.7 Classification report of AdaBoost for "No Torture".

	Precision	Recall	F1-score	Support
0	0.93	0.83	0.88	76
1	0.57	0.77	0.65	22
Micro Average	0.82	0.82	0.82	98
Macro Average	0.75	0.80	0.76	98
Weighted Average	0.85	0.82	0.83	98

Table 4.8 Classification report of Decision Tree for "No Unfair Detainment".

	Precision	Recall	F1-score	Support
0	0.91	0.97	0.93	89
1	0.00	0.00	0.00	9
Micro Average	0.88	0.88	0.88	98
Macro Average	0.45	0.48	0.47	98
Weighted Average	0.82	0.88	0.85	98

To determine the best classifier, we are using three different measurement techniques:

1. **Precision** - It is the number of positive predictions and divided by the total positive class values predicted. It is the measure of a classifier's exactness.

$$Precision = \frac{Relevant\ articles\ \cap Retrieved\ articles}{Retrieved\ articles} \tag{4.1}$$

2. **Recall** - It is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate. It is the measure of classifier's completeness.

$$Recall = \frac{Relevant\ articles\ \cap Retrieved\ articles}{Relevant\ articles} \tag{4.2}$$

3. **F1 Score** - It is the balance between the precision and recall and is calculated by:

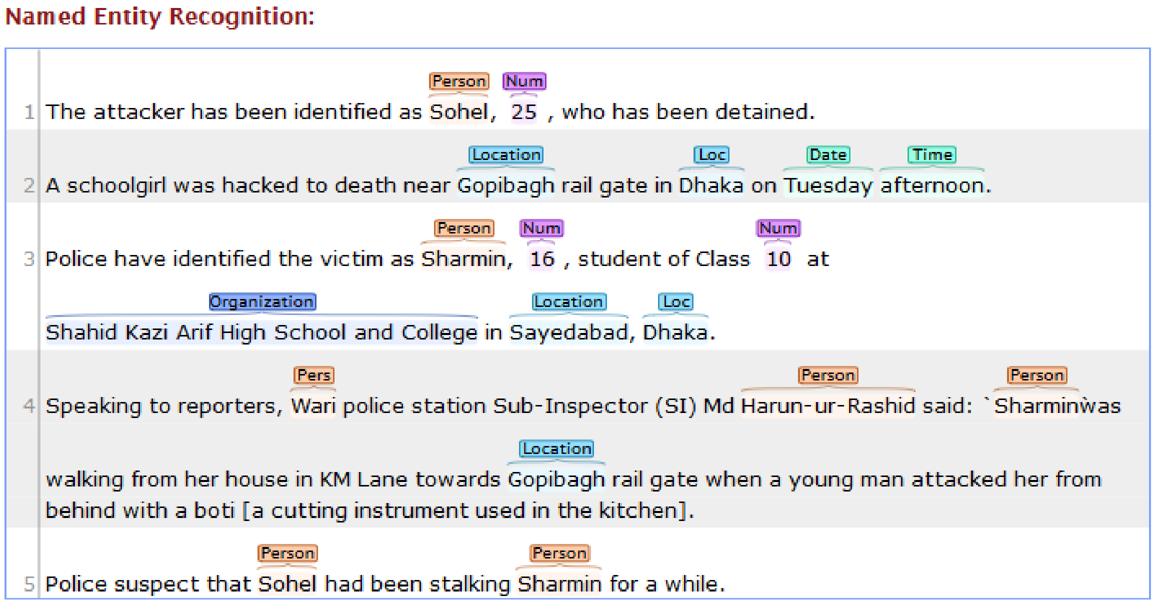
$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.3}$$

As we have experimented with all these classification algorithms, it has been found that each algorithm performs differently for each category. We have concluded that **SVM** achieves the best accuracy for the purpose of our research.

4.3.3 Location Extraction

Stanford's CoreNLP has been used to extract the location specified in the articles by running Named Entity Recognition model. The system first tokenizes the whole article into sub-components and is represented as a **CoreDocument** object. This object is then passed to a co-reference extractor and annotator to determine different entity names.

The figure below provides a visual overview of how the model extracts different entity names:



Named Entity Recognition:

1 The attacker has been identified as Sohel, 25 , who has been detained.

2 A schoolgirl was hacked to death near Gopibagh rail gate in Dhaka on Tuesday afternoon.

3 Police have identified the victim as Sharmin, 16 , student of Class 10 at Shahid Kazi Arif High School and College in Sayedabad, Dhaka.

4 Speaking to reporters, Wari police station Sub-Inspector (SI) Md Harun-ur-Rashid said: `Sharminẁas walking from her house in KM Lane towards Gopibagh rail gate when a young man attacked her from behind with a boti [a cutting instrument used in the kitchen].

5 Police suspect that Sohel had been stalking Sharmin for a while.

Fig. 4.6 Extracting location using NER.

An article may contain multiple location name. In this situation, it can be done two things:

1. Take the location with the highest number of occurrences within the article.
2. Take the location which resides in the same sentence as the main verb of the article.

4.3.4 Visualization of Incidents

Ultimately, after successfully extracting the geographical location of the incidents using Stanford's CoreNLP, we have represented the data in a heat map, to visualize the state of occurrences though out the country.

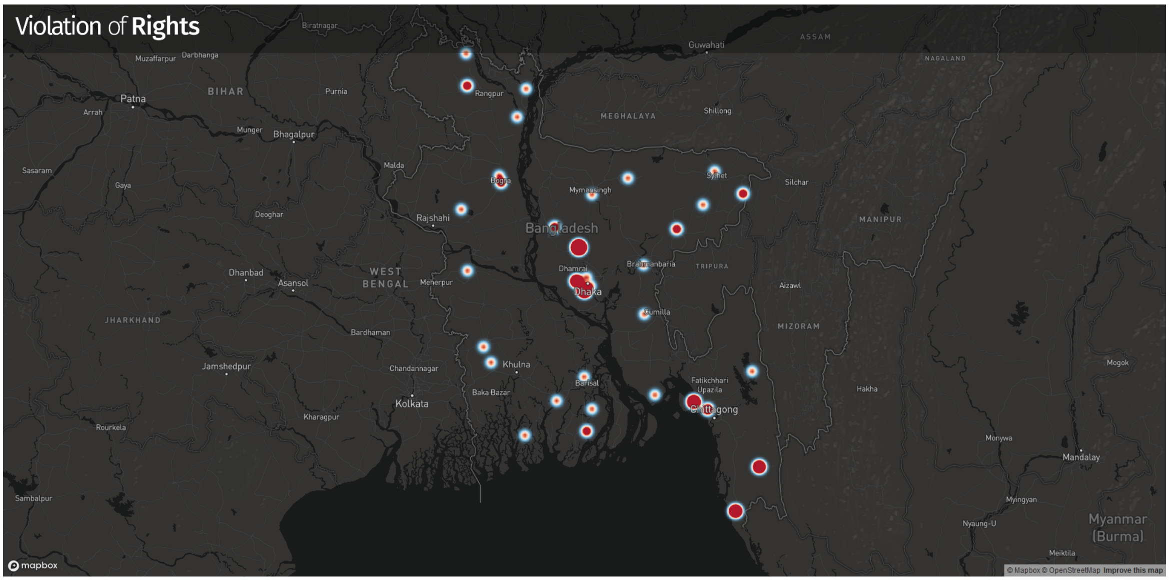
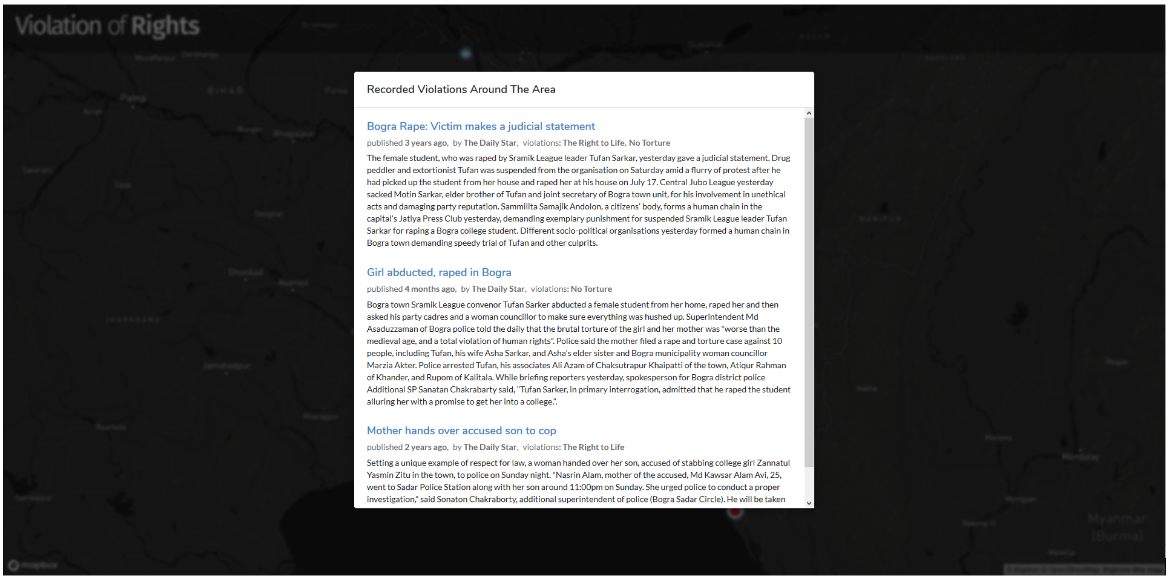


Fig. 4.7 What the visitors see when they visit the website.



Recorded Violations Around The Area

Bogra Rape: Victim makes a judicial statement

published **3 years ago**, by **The Daily Star**, violations: **The Right to Life**, **No Torture**

The female student, who was raped by Sramik League leader Tufan Sarkar, yesterday gave a judicial statement. Drug peddler and extortionist Tufan was suspended from the organisation on Saturday amid a flurry of protest after he had picked up the student from her house and raped her at his house on July 17. Central Jubo League yesterday sacked Motin Sarkar, elder brother of Tufan and joint secretary of Bogra town unit, for his involvement in unethical acts and damaging party reputation. Sammilita Samajik Andolon, a citizens' body, forms a human chain in the capital's Jatiya Press Club yesterday, demanding exemplary punishment for suspended Sramik League leader Tufan Sarkar for raping a Bogra college student. Different socio-political organisations yesterday formed a human chain in Bogra town demanding speedy trial of Tufan and other culprits.

Girl abducted, raped in Bogra

published **4 months ago**, by **The Daily Star**, violations: **No Torture**

Bogra town Sramik League convenor Tufan Sarker abducted a female student from her home, raped her and then asked his party cadres and a woman councillor to make sure everything was hushed up. Superintendent Md Asaduzzaman of Bogra police told the daily that the brutal torture of the girl and her mother was "worse than the medieval age, and a total violation of human rights". Police said the mother filed a rape and torture case against 10 people, including Tufan, his wife Asha Sarkar, and Asha's elder sister and Bogra municipality woman councillor Marzia Akter. Police arrested Tufan, his associates Ali Azam of Chaksutrapur Khaipatti of the town, Atiqur Rahman of Khander, and Rupom of Kalitala. While briefing reporters yesterday, spokesperson for Bogra district police Additional SP Sanatan Chakrabarty said, "Tufan Sarker, in primary interrogation, admitted that he raped the student alluring her with a promise to get her into a college.".

Mother hands over accused son to cop

published **2 years ago**, by **The Daily Star**, violations: **The Right to Life**

Setting a unique example of respect for law, a woman handed over her son, accused of stabbing college girl Zannatul Yasmin Zitu in the town, to police on Sunday night. "Nasrin Alam, mother of the accused, Md Kawsar Alam Avi, 25, went to Sadar Police Station along with her son around 11:00pm on Sunday. She urged police to conduct a proper investigation," said Sonaton Chakraborty, additional superintendent of police (Bogra Sadar Circle). He will be taken

Fig. 4.8 What the visitors see when they click on a point of the map.

4.4 Data Analysis

Using the dataset, a simple analysis of frequency of incidents, location of the incidents and date and time of the occurrences are shown below.

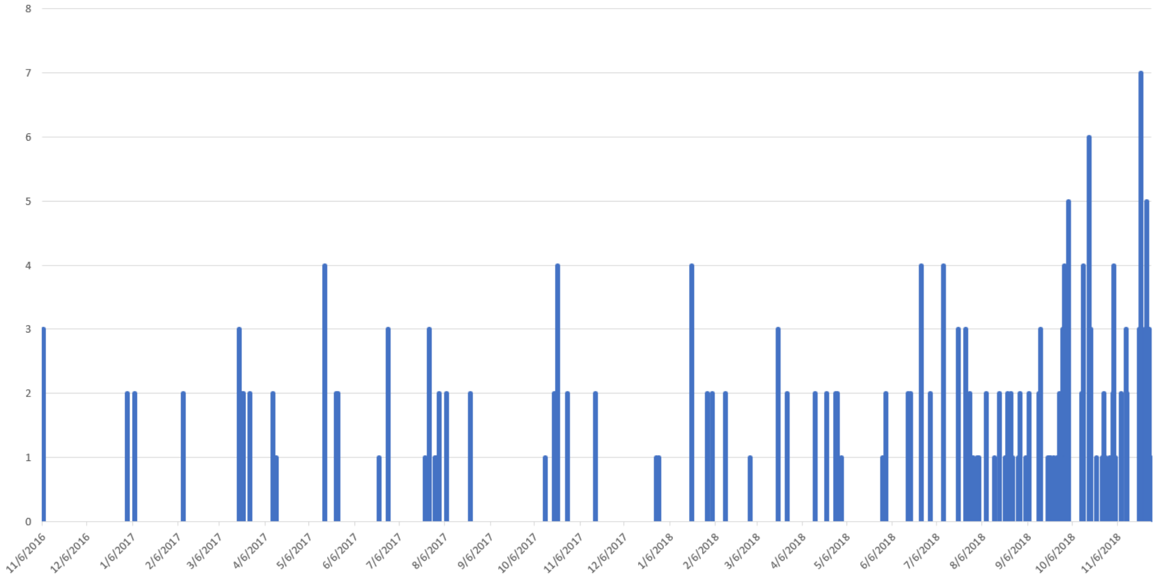


Fig. 4.9 Number of incidents recorded from 2016 to 2018.

We can see that the number of occurrences are increasing. From this, we can infer two things:

1. The number of violations are increasing.
2. The incidents of violation are being recorded by the news sources more frequently.

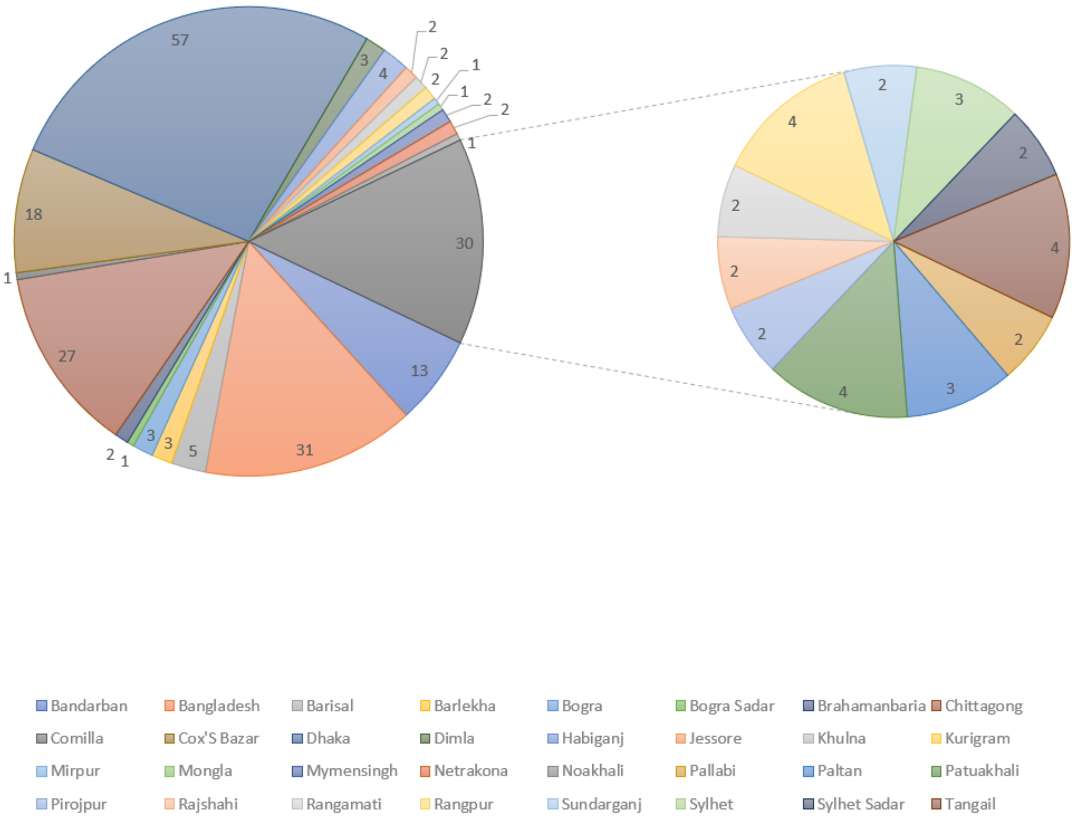


Fig. 4.10 Number of incidents recorded in each area.

In this analysis, it has been found that remote areas such as Bandarban has higher rates of violation compared to urban areas. This information is a valuable resource for the law enforcement agencies to focus on specific areas. We can also infer that, since these remote areas have people who are deprived of basic education, thus people are more prone to cause these kinds of violations.

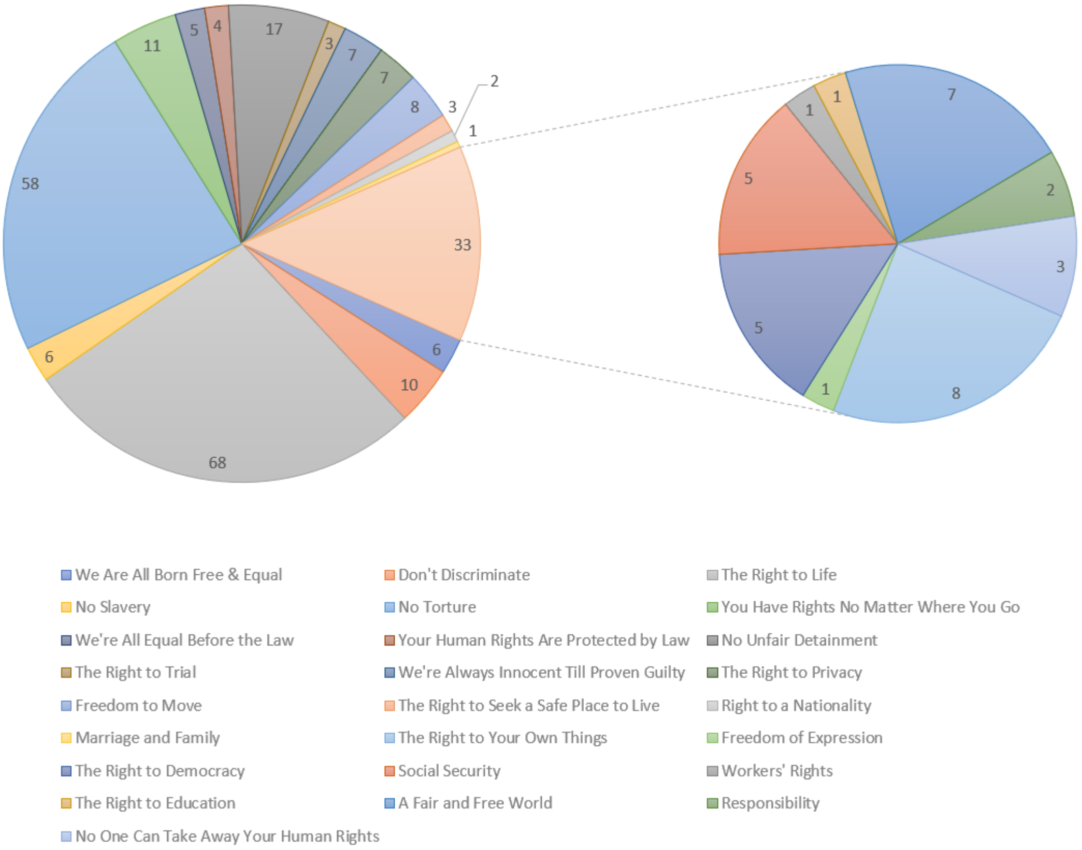


Fig. 4.11 Number of violations per human right.

Among all the human rights, the most violated ones are "The right to life" and "We are all born free and equal". Since enough precautions are not taken, people are getting away with the crimes they are committing. Thus, these rights are among the top.

Chapter 5

Conclusion

This research finds out about all the events that involve violation of Human Rights in Bangladesh using news articles from The Daily Star, Prothom Alo and Dhaka Tribune. We have successfully been able to classify the articles according to Human Rights along with extracting the location where the event has taken place. Ultimately, presenting the occurrences in a map for visualization.

Using multiple classifying algorithms, it has been analyzed that the best way to classify incidents is to use SVM, according to the performance metrics.

Throughout the research, we have been able to find the different types of human right violations in different areas of Bangladesh, recording events that were unnoticed for public consumption. One of our goals is to make the data accessible to the Governments and the NGOs to take proper measure against the violations. This research contributes in the field of work that strives to identify and prevent actions against basic Human Rights to make the world a better place.

5.1 Limitations

We have faced some bottlenecks and intricacies despite our model being successful, during the research. The problem lies in the datasets being imbalanced and small in size which means that there are very few incidents for some of the human right points than the other, and classifiers do not have enough data to accurately classify the articles according to their category.

Due to lack of enough data, we are unable to carry out Regression Analysis which would enable us to project Frequency against Time of occurrences and Frequency versus Location of occurrences to subsequently predict the likelihood of frequency of Human Rights violation in a specific year and its probable location(s).

Lastly, articles containing ambiguous information is another limitation of our model. The model is unable to identify fake news, so if an article contains misinformation, it would accept it as if the article falls within the category of the rights violation.

5.2 Future Works

Our model could be improved and utilized in various causes. To begin with, with enough dataset, the usability of predicting machine learning models could be made available. Prediction of frequency of violations against time could be a great resource to the Government and organizations fighting against the acts.

In addition to that, our data bank can be used to portray what Bangladeshi laws have to say about a particular violation, raise public awareness and educate people about the magnitude of the crime.

Furthermore, our research has solely relied on news articles from three different news agencies. More news websites can be added to enrich the database, providing more information for the algorithms to work on. Also, this research was done taking only the news articles from Bangladesh. It could be extended to cover other regions of the World as well.

Lastly, the detection of violation of Human Rights can also be done over the domain of social media content which contains a lot of information being unnoticed of. Our model could be modified and trained for the social media contents to analyze the infringements.

References

[1] Absar, K., Gebru, E., Linnell, P., Montal, F., and O'Donnell, K. (2015). Event-based media monitoring methodology for human rights watch.

[2] Baker, L. D. and McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103. ACM.

[3] Cai, D., Yu, S., Wen, J.-R., and Ma, W.-Y. (2003). Extracting content structure for web pages based on visual representation. In *Asia-Pacific Web Conference*, pages 406–417. Springer.

[4] Chandler, D. (2001). The road to military humanitarianism: How the human rights ngos shaped a new humanitarian agenda. *Hum. Rts. Q.*, 23:678.

[5] Chen, F. and Neill, D. B. (2014). Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1166–1175. ACM.

[6] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). A framework and graphical development environment for robust nlp tools and applications. In *ACL*, pages 168–175.

[7] Davenport, C. and Armstrong, D. A. (2004). Democracy and the violation of human rights: A statistical analysis from 1976 to 1996. *American Journal of Political Science*, 48(3):538–554.

[8] Eriksson, O. and Rydkvist, E. (2015). An in-depth analysis of dynamically rendered vector-based maps with webgl using mapbox gl js.

[9] Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3):131–163.

[10] Han, E.-H. S., Karypis, G., and Kumar, V. (2001). Text categorization using weight adjusted k-nearest neighbor classification. In *Pacific-asia conference on knowledge discovery and data mining*, pages 53–65. Springer.

[11] Hovy, E. and Lin, C.-Y. (1998). Automated text summarization and the summarist system. In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*, pages 197–214. Association for Computational Linguistics.

[12] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

[13] Kim, Y.-H., Hahn, S.-Y., and Zhang, B.-T. (2000). Text filtering by boosting naive bayes classifiers. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 168–175. ACM.

[14] Laender, A. H., Ribeiro-Neto, B. A., Da Silva, A. S., and Teixeira, J. S. (2002). A brief survey of web data extraction tools. *ACM Sigmod Record*, 31(2):84–93.

[15] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

[16] Marrin, C. (2011). Webgl specification. *Khronos WebGL Working Group*.

[17] Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

[18] Morsink, J. (1999). *The Universal Declaration of Human Rights: origins, drafting, and intent*. university of Pennsylvania Press.

[19] Patel, S. (2017). Chapter 2 : Svm (support vector machine) - theory – machine learning 101 – medium. https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72.

[20] Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142.

[21] Sidelnikov, G. (2018). Webgl pipeline. http://www.webgltutorials.org/pipeline.html.

[22] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958.

[23] Tan, A.-H. et al. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, pages 65–70. sn.

[24] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69–90.