মেট্রোপলিটন ইউনিভার্সিটি
Metropolitan University

A thesis
Submitted to
The Department of Computer Science and Engineering, Metropolitan University,
in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

CLASSIFICATION AND ANALYSIS OF ROAD ACCIDENT NEWS ARTICLES USING MACHINE LEARNING

Submitted by:
Nusrat Farzana (151-115-106)
Foysol Ahmed Shuvo (151-115-120)
Md. Nafiul Adnan Chowdhury (151-115-123)

Under the supervision of:
Sakhawat Hossain Saimon
Lecturer,
Department of CSE,
Metropolitan University.

January, 2019

Declaration

We hereby declare that the thesis is our original work and it has been written by us in its entirety. We have duly acknowledged all the sources of information which have been used in the thesis.

The thesis has also not been submitted for any degree in any university previously.

---	---	---
Nusrat Farzana	Foysol Ahmed Shuvo	Md. Nafiul Adnan Chowdhury

Recommendation Letter from Thesis Supervisor

These students, Nusrat Farzana, Foysol Ahmed Shuvo, Md. Nafiul Adnan Chowdhury whose thesis entitled “Classification And Analysis of Road Accident News Articles Using Machine Learning”, is under my supervision and agrees to submit for examination.

Advisor:

Sakhawat Hossain Saimon

Lecturer,Department of CSE,

Metropolitan University.

Date: _____

…….

Advisor

Qualification From of Bachelor Degree

Students Name: Nusrat Farzana, Foysol Ahmed Shuvo, Md. Nafiul Adnan Chowdhury

Thesis Title: Classification And Analysis of Road Accident News Articles Using Machine Learning

This is to certify that the thesis submitted by the student named above in January, 2019. It is qualified and approved by the Thesis Examination Committee.

---	---	---
Head of the Dept.	Chairman, Thesis Committee	Supervisor

Acknowledgment

We are thankful to our supervisor “Sakhawat Hossain Saimon” for his support and direction to our working progress. We also thank Prothom-alo’s website from where we got data for our thesis work and other websites for helping us to complete our thesis.

Abstract

In Bangladesh, Road accidents are occurring almost on daily basis and taking precious lives, wealth and leaving people severely injured for the rest of their lives. Although a lot of steps are being taken to get rid of it, the situation is not improved. In this research paper, as accident data source, we have used Prothom-Alo online newspaper's accident news of all over the country. We tried to find the road accident occurrence rate and statistics of all districts of Bangladesh and plotted them in Heatmap. We showed the vehicles that are responsible for most of the road accidents as well. To detect road accident news and find the accident location and vehicles, we used machine learning algorithms with pattern matching. Also, provided a comparison of performance between three machine learning algorithms – Naive Bayes, Cosine Similarity and Support Vector Machine (SVM).

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Road accident became a major issue nowadays. In the year 2009 in Bangladesh, 3381 road accidents were recorded where 2958 people died and 2686 were injured[30]. The causalities decreased a little bit later. According to World Bank data, deaths per 100,000 people in Bangladesh dropped from 14.1 in 2010 to 12.8 in 2015 [31]. But in 2017, 2297 people died and 5480 was injured only in the period of 6 months (from January to June)[32]. According to a study in Bangladesh fatalities per 100,000 population is 13.6% and in our neighboring country, India it is 16.6%[33]. Though the rate is high in India, if we compare the population we can realize how higher it is in our country. There are approximately 316,000 deaths each year because of road accidents that occur in the South-East Asia Region[33].

So, we thought if there is any way we can help the government of Bangladesh and the general people on this delicate matter.

There are a lot of causes to occur a road accident. It can be occurred because of rapid driving, the tendency of sudden overtaking, unfit vehicles, the structure of the road, unaware pedestrian or many other issues. So, it would be helpful if we know the accident rate of any specific area that we can observe why accidents are happening there, by analyzing the rate. The government might take a look at these areas to find out the reasons and also travelers and drivers also would be more conscious of their choice of area of travel.

1.1 Previous Work

The amount of research work in Bengali Natural Language Processing is not as rich as English [18]. Therefore, there was not enough research work has been done on road accidents of Bangladesh. Although, there are some works to measure the injury caused by road accidents, [19] the data was the GES automobile accident data from 1995 to 2000.

To mining traffic data of N5 National highway in Bangladesh [20], there are some works. But they worked only on the basis of a specific national highway using 892 road accidents data. There is another work to detect the road accident and the anomaly detection from roadside video data [21], but they did not use any machine learning approach and also the data was insufficient as they used only the roadside video data where, in Bangladesh, there are not enough roads having roadside video cameras.

There was a relevant work where statistics of road accidents along with vehicles [22] was shown. But they showed the comparison between Dhaka with the whole Bangladesh only. We can see a diagram which was used by them [22].

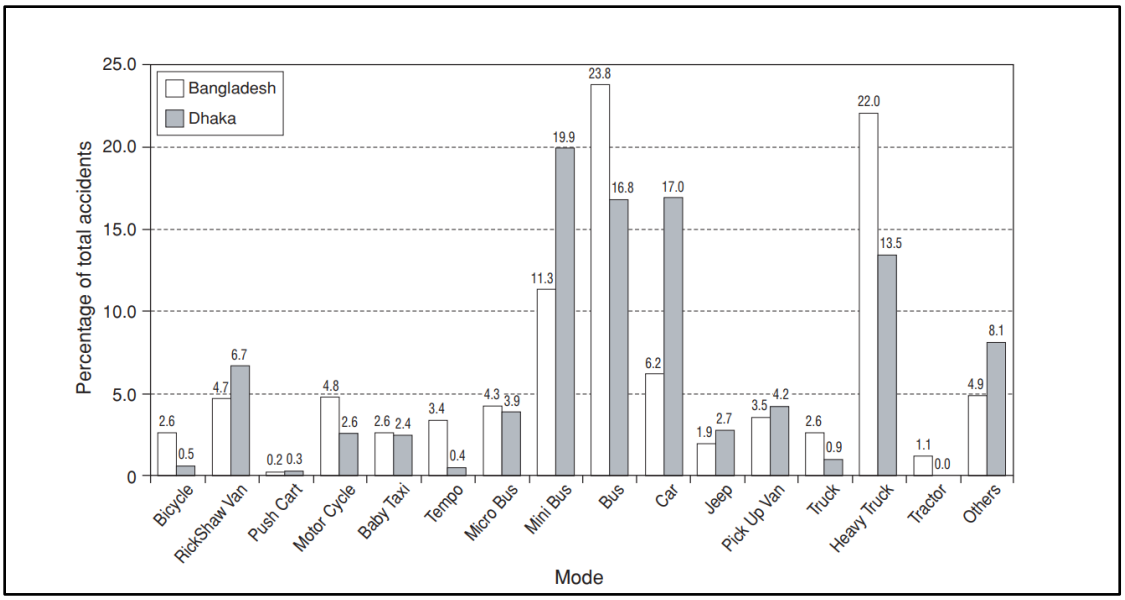


Figure 1: Comparison between Dhaka and Bangladesh (without Dhaka) road accidents

Another work has been done to analyze road accident data [23] which was only for Dhaka city, and the accident data was collected from Dhaka metropolitan police.

The most relevant work to our work was another research paper where they worked on predicting the crime occurrence in all over Bangladesh using machine learning approach [4].

We needed a huge amount of data on road accidents from all over the country. So, we picked the most popular online newspaper- 'Prothom-Alo'. We crawled all the accident news from the newspaper. Using machine learning algorithms, we detected the road accident news from there and we tried to come up with an approach where we can see all the statistics of road accidents from all the districts of Bangladesh, detect the vehicles that are mostly involved in road accidents. Also, we tried to show a comparison of the machine learning algorithms we have used.

1.2 Our Goals

Our goal is to provide a statistic of accidents occurring over years and plot them in the map also. To get the road accident occurrence of different districts has been collected after crawling different news from Prothom-Alo online newspaper. This system also provides a map where different road accidents pointed according to their occurrence zone. These might be helpful for general people, tourists to decide their path and metro police to aware of where road accidents are occurring most frequently. So, the accident rate can be decreased and we can have a safer journey.

CHAPTER 2

Background Study

2.1 Web crawler

A web crawler is also known as Spider or robot, it is a program that starts with a list of URLs called the Seeds which it needs to visit. Crawler downloads the web pages associated with these seeds extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks[1]. It is an important component of the search engine because it is used to collect the corpus of web pages indexed by the engine [2]. Crawlers are used for many other purposes. For example, some websites use them to update their web content, and they can validate hyperlinks and HTML code also used for web scraping [1].

Crawlers consume resources when they visit systems and problems like schedule, load and politeness come when it access a large number of pages. For this reason, search engines struggled to give relevant results in the early years before 2000. But today it is done instantly[25].

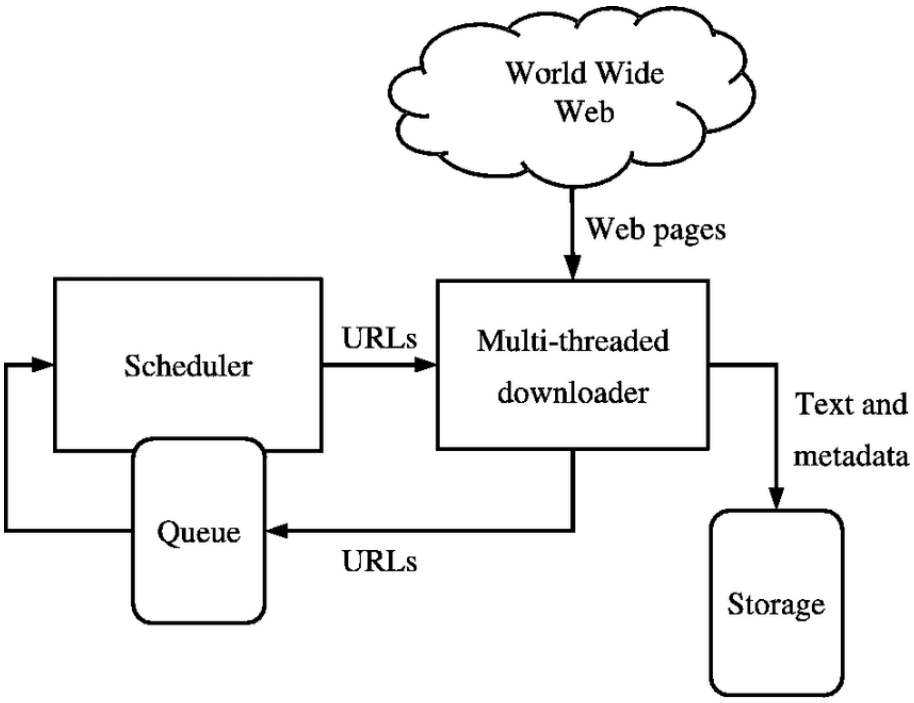


Figure 2: Web Crawler

2.2 Parsing

Parsing is involved with extracting hyperlinks or other complex process as tidying up HTML content, it may also involve steps as converting extracted hyperlinks to a canonical form, remove stop words from page's content and stem keywords[4]. Components of pursing which we used in our thesis such as stop words, keywords etc. are described next.

2.3 Stop words

After parsing content from a web page it is always helpful to remove frequently or most common words which we call stop words such as "to", "the", "a", "it", "of" etc.[24]. In the Bengali language, there is also some words which are known as stop words such as "এটা" , "হয়" , " যায়" , " ইহা" etc.[4]. The process where we cut stop words is referred as to stop word listing[4].

There is another problem. "Try" and "Tried" means the same thing but they are not in the same tense. So computer won't understand that they are same in meaning. So tried is converted to try. This is called stemming process. For Bengali "করছে" and "করছিল" are in the same meaning but different tense which is brought to "করছে".

2.4 Keyword Extraction

Key phrases, key terms, key segments or just keywords are the terminology which is used for defining the terms that represent the most relevant information contained by the document[5]. It is used for automatic identification of terms that best describe the subject of the document, document retrieval, document clustering, web page retrieval, text mining and so on[4]. Algorithms such as TF-IDF use this.

2.5 Clustering

Clustering is a process that groups similar objects together in a subset which is called a cluster. It is an unsupervised learning technique that helps us to find similarities between objects like data point and group them together and object in a group is more similar to the objects of same groups than objects in other groups [6]. Therefore, A cluster is a collection of objects which are similar with each other and dissimilar to the objects belongs to other clusters.

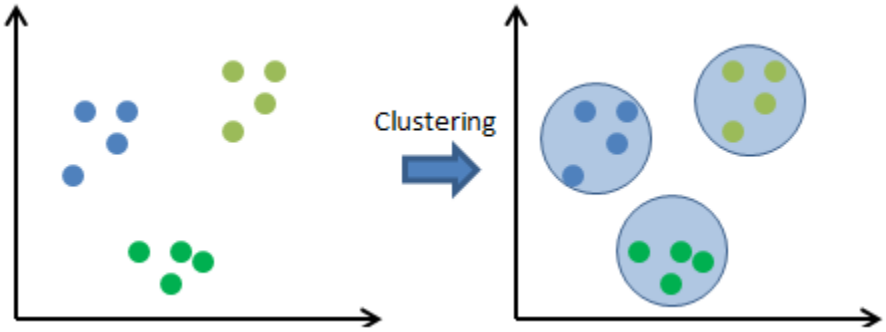


Figure 3: Clustering

In Figure 3, we can see that on the left side there are some data points scattered and there is also some data which is similarly based on their color. So if we cluster them blue color data will be in one cluster green in another, which we can see on the right side of the picture.

2.4 K-means Clustering

K-means is one of the simplest and easy clustering techniques than all other clustering algorithms. It is popular for cluster analysis in data mining. This clustering technique partition n objects into k clusters where each object belongs to the cluster with the nearest mean, initially we take k random data points from our data to make k clusters, then from these k data points we measure the distance of every data points using the Euclidian equation[8]:

$$\boldsymbol{d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}$$

The point that has the shortest distance with a cluster goes to that cluster. After that step, we measure the mean of each cluster to find new coordinates of clusters. Iteration continues like that until results of two iterations are almost the same and only then we find our data clustered into k clusters[8].

2.5 Naive Bayes Classifier

It is one of the simplest techniques for constructing classifiers. Naive Bayes classifiers are a family of simple probabilistic classifiers which uses Bayes theorem with strong (naive) independence assumptions[9]. In simple word, Naive Bayes classifier finds the probability of being of an object in clusters based on given train dataset and the class that has high probability. It adds this object to that cluster. For example, if we give a train data set which says apple is red and mango is green then whenever it will find apple it will put that in the red color group.

The Naive Bayes classifier has been widely used as it has simplicity in both the training and classifying stage[10]. This algorithm is effective in text classification, medical diagnosis etc. [9].

2.6 Cosine Similarity

Cosine similarity measures similarity between two non-zero vectors of an inner product space which measures the cosine of the angle between them and it compares between documents on a normalized space because of it considerate not only the magnitude of each word count (TF-IDF) of each document also the angle between the documents[11]. We just have to solve cosine similarity formula which is[11]:

$$\boldsymbol{Similarity} = \boldsymbol{cos\,\theta} = \frac{\boldsymbol{A}\,.\,\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} = \frac{\sum_{i=1}^{n} \boldsymbol{A_i B_i}}{\sqrt{\sum_{i=1}^{n} \boldsymbol{A_i^2}}\sqrt{\sum_{i=1}^{n} \boldsymbol{B_i^2}}}$$

2.7 TF-IDF

Term frequency-inverse document frequency which means TF-IDF is a numerical statistic that shows how important a word is to a document[4]. The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others[12]. TF–IDF is one of the most popular term-weighting schemes today [12].

2.8 Kth Nearest Neighbor

Kth nearest neighbor in short K-**NN** is a non-parametric method which is used for classifying an unclassified data point based on a set of classified data near the test data point[13].

In K-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor[14].

KNN is the simplest of all the algorithms of machine learning[25].

2.9 Jaccard Similarity

Jaccard Similarity is a statistic used for comparing the similarity between sets. The Jaccard coefficient measures the similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets[15]:

$$\boldsymbol{J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}}$$

Another term which comes with Jaccard similarity is Jaccard Distance, which shows dissimilarity between two sets[15]. We can calculate that by subtracting the value of Jaccard Similarity coefficient from 1. So the formula is[15]:

$$\boldsymbol{D(X,Y) = 1 - J(X,Y)}$$

2.10 Hidden Mark Model

Hidden Mark Model in short HMM is a statistical Markov model. In HMM the system which is being modeled is assumed to be a Markov process with unobserved states. In a hidden Markov model, the state is not directly visible like Markov Model[6]. Its output is dependent on states and only these states are visible. Each state has a probability distribution over the possible output tokens[6]. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

HMM is considered as the simplest dynamic Bayesian network[4].

2.11 Support Vector Machine

The Support Vector Machine (SVM) was first proposed by Vapnik and got a high degree of interest to the machine learning researcher [29].

SVMs are a set of related supervised learning methods used for classification and regression [29].

An SVM Classifier is a machine learning algorithm that analyzes data used for classification and regression analysis. If there is given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier[28].

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

2.12 Neural Network

Every human brain has some special cells called neurons. And they communicate with each other using electrical impulses. Like Dendrites receive input signals based on inputs and fire an output signal via an axon or something like that. Some computer scientists had the idea that is modeled after this system of neural connections and they called their idea "Neural Networks"[27].

Neural networking has several types of learning like, Supervised Learning, Unsupervised Learning, Reinforcement Learning and etc.

Today, there are so many uses of the neural network. Here are some standard uses of neural networks. For example[26],

- Pattern Recognition
- Time Series Prediction
- Signal processing
- Self-driving cars
- Soft Sensors
- Anomaly Detection

2.13 Precision and Recall

Precision which is also called Positive Predictive value is the fraction of relevant instances among the retrieved instances[16].

Recall which is also known as Sensitivity is the fraction of relevant instances that have been retrieved over the total amount of relevant instances[16].

They both are used in pattern recognition information retrieval and binary classification.

To measure precision and recall we use the following equations[16]:

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive} + \boldsymbol{False\ Positive}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive} + \boldsymbol{False\ Negative}}$$

Here,

True Positive is an outcome where the model correctly predicts the positive class

True Negative is an outcome where the model correctly predicts the negative class.

False Positive is an outcome where the model incorrectly predicts the positive class

False Negative is an outcome where the model incorrectly predicts the negative class.

For examples,

If the model predicts Nike shoe as Nike then it is True Positive,

Predicts non-Nike shoe as non-Nike then True Negative,

Predicts non-Nike shoe as Nike then False Positive,

Predicts Nike shoe as non-Nike then False Negative.

2.14 Accuracy

To measure a test's accuracy we use the following equation[16]:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}$$

2.15 F-Beta measure

In a statistical analysis of binary classification, the F Beta-measure is a measure of a test's accuracy in the simple word we use that to find which system is suitable.

It uses precision and recalls along with a constant called beta which value can be between 0 and 1 inclusive. F Beta measure reaches its best value at Beta=1 (perfect precision and recall) and worst at Beta=0, the higher the value of beta, the higher the importance to precision[17]. The Greater the value of F Beta the Higher suitable the system is. The Equation of F Beta is given bellow[17]:

$$\boldsymbol{F\beta} = \frac{\mathbf{1}}{\boldsymbol{\beta}\times\left(\frac{\mathbf{1}}{\boldsymbol{Precision}}\right)+(\mathbf{1}-\boldsymbol{\beta})\times\frac{\mathbf{1}}{\boldsymbol{Recall}}}$$

CHAPTER 3

Methodology

Our methodology consists of different levels. At the very first level, we crawled data and modified. Then we categorized the data using machine learning algorithms and went for further processing as extracting location vehicles creating heatmap and result analysis.

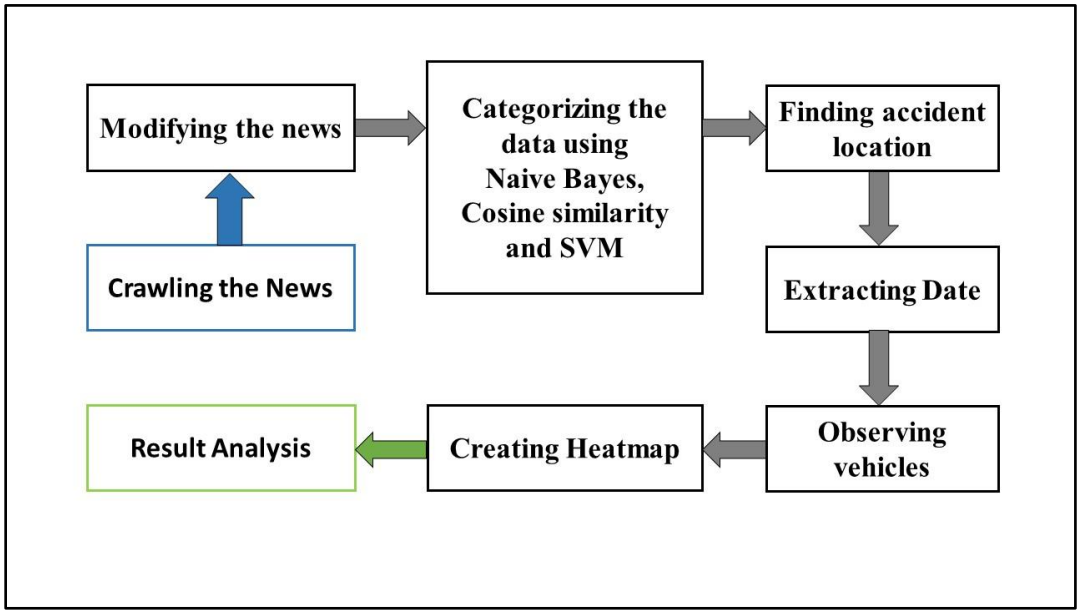


Figure 4: Implementation procedure

3.1 News Crawling

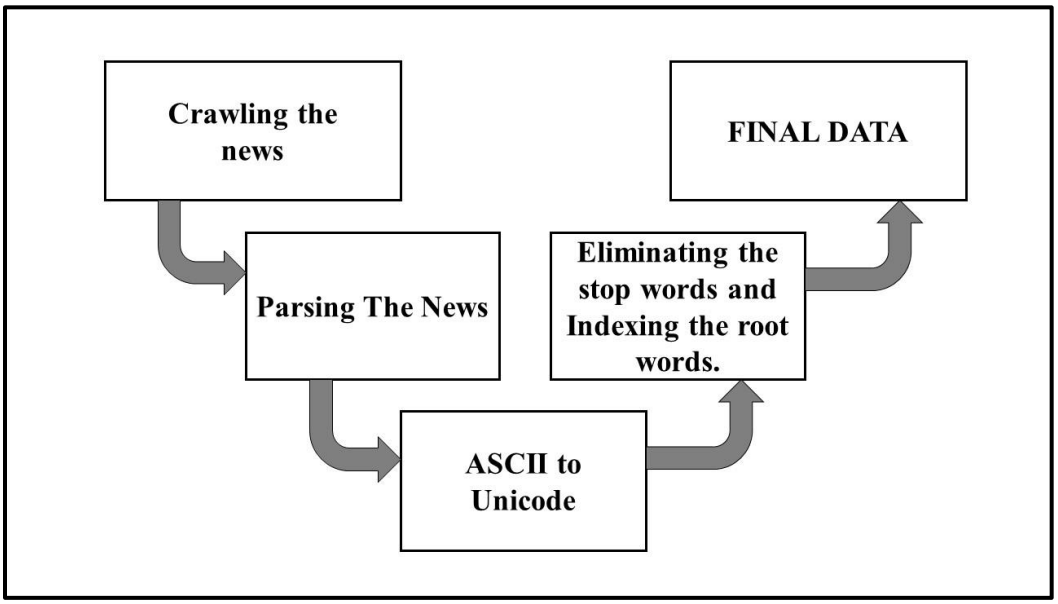


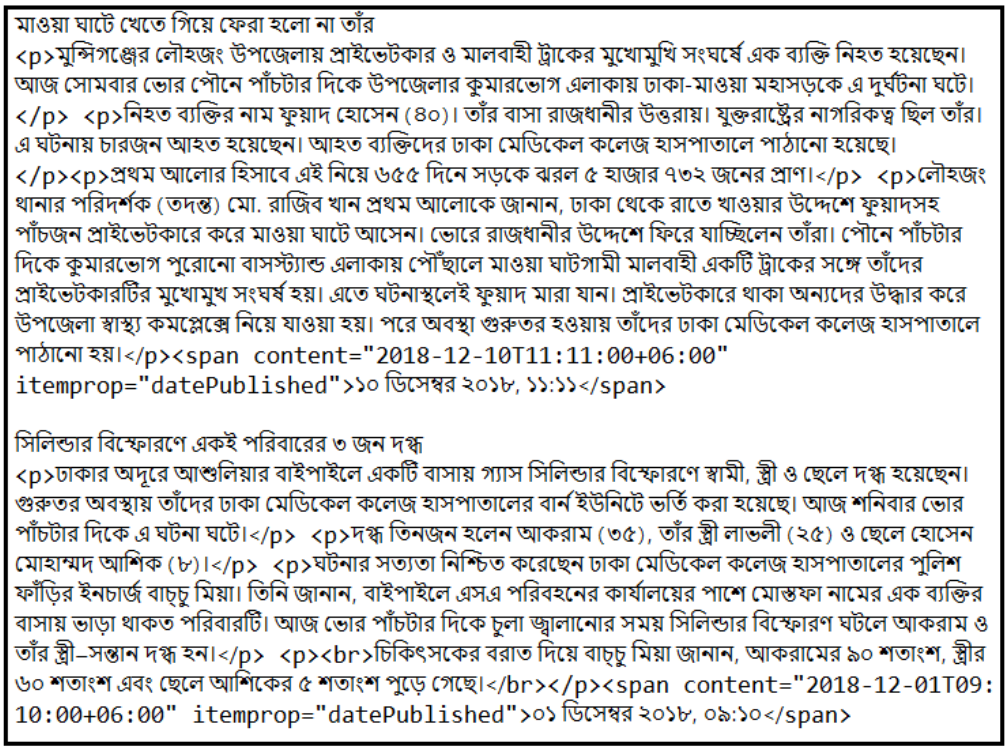
Figure 5: Data parsing procedure

For crawling the news which is our main data, we used Prothom-Alo newspaper's website and we crawled all the news of different types of accident using python code. Crawled news containing the headline, news and publishing date of the news.

In our code, we imported the "requests" and "BeautifulSoup" libraries of python. The "requests" used to get the whole text from the news from the provided URL. The "BeutifulSoup" templates did all the other things such as finding the headlines, finding all the paragraphs from the news and the date of the news.

We separated all the attributes from the news page and wrote these in a file which was our crawled data. We found that in the News paragraph, there were some images of that particular event included to the News too. As we crawled all the paragraphs from a News, the tags for the images were also crawled.

We figured out that in each page of the Prothom-Alo's website, there was exactly 20 news which was relevant to Accident tags. Other news was some suggestions for different types of News. such as- Politics, Entertainment, Sports, etc. As we needed to work with accidents news only, we crawled exactly 20 news from each of the pages which were only the news of different types of accident.



মাওয়া ঘাটে খেতে গিয়ে ফেরা হলো না তাঁর
মুন্সিগঞ্জের লৌহজং উপজেলায় প্রাইভেটকার ও মালবাহী ট্রাকের মুখোমুখি সংঘর্ষে এক ব্যক্তি নিহত হয়েছেন। আজ সোমবার ভোর পৌনে পাঁচটার দিকে উপজেলার কুমারভোগ এলাকায় ঢাকা-মাওয়া মহাসড়কে এ দুর্ঘটনা ঘটে। নিহত ব্যক্তির নাম ফুয়াদ হোসেন (৪০)। তাঁর বাসা রাজধানীর উত্তরায়। যুক্তরাষ্ট্রের নাগরিকত্ব ছিল তাঁর। এ ঘটনায় চারজন আহত হয়েছেন। আহত ব্যক্তিদের ঢাকা মেডিকেল কলেজ হাসপাতালে পাঠানো হয়েছে। প্রথম আলোর হিসাবে এই নিয়ে ৬৫৫ দিনে সড়কে ঝরল ৫ হাজার ৭৩২ জনের প্রাণ। লৌহজং থানার পরিদর্শক (তদন্ত) মো. রাজিব খান প্রথম আলোকে জানান, ঢাকা থেকে রাতে খাওয়ার উদ্দেশে ফুয়াদসহ পাঁচজন প্রাইভেটকারে করে মাওয়া ঘাটে আসেন। ভোরে রাজধানীর উদ্দেশে ফিরে যাচ্ছিলেন তাঁরা। পৌনে পাঁচটার দিকে কুমারভোগ পুরোনো বাসস্ট্যান্ড এলাকায় পৌঁছালে মাওয়া ঘাটগামী মালবাহী একটি ট্রাকের সঙ্গে তাঁদের প্রাইভেটকারটির মুখোমুখ সংঘর্ষ হয়। এতে ঘটনাস্থলেই ফুয়াদ মারা যান। প্রাইভেটকারে থাকা অন্যদের উদ্ধার করে উপজেলা স্বাস্থ্য কমপ্লেক্সে নিয়ে যাওয়া হয়। পরে অবস্থা গুরুতর হওয়ায় তাঁদের ঢাকা মেডিকেল কলেজ হাসপাতালে পাঠানো হয়। ১০ ডিসেম্বর ২০১৮, ১১:১১

সিলিন্ডার বিস্ফোরণে একই পরিবারের ৩ জন দগ্ধ
ঢাকার অদূরে আশুলিয়ার বাইপাইলে একটি বাসায় গ্যাস সিলিন্ডার বিস্ফোরণে স্বামী, স্ত্রী ও ছেলে দগ্ধ হয়েছেন। গুরুতর অবস্থায় তাঁদের ঢাকা মেডিকেল কলেজ হাসপাতালের বার্ন ইউনিটে ভর্তি করা হয়েছে। আজ শনিবার ভোর পাঁচটার দিকে এ ঘটনা ঘটে। দগ্ধ তিনজন হলেন আকরাম (৩৫), তাঁর স্ত্রী লাভলী (২৫) ও ছেলে হোসেন মোহাম্মদ আশিক (৮)। ঘটনার সত্যতা নিশ্চিত করেছেন ঢাকা মেডিকেল কলেজ হাসপাতালের পুলিশ ফাঁড়ির ইনচার্জ বাচ্চু মিয়া। তিনি জানান, বাইপাইলে এসএ পরিবহনের কার্যালয়ের পাশে মোস্তফা নামের এক ব্যক্তির বাসায় ভাড়া থাকত পরিবারটি। আজ ভোর পাঁচটার দিকে চুলা জ্বালানোর সময় সিলিন্ডার বিস্ফোরণ ঘটলে আকরাম ও তাঁর স্ত্রী–সন্তান দগ্ধ হন।
চিকিৎসকের বরাত দিয়ে বাচ্চু মিয়া জানান, আকরামের ৯০ শতাংশ, স্ত্রীর ৬০ শতাংশ এবং ছেলে আশিকের ৫ শতাংশ পুড়ে গেছে।</br> ০১ ডিসেম্বর ২০১৮, ০৯:১০

Figure 6: Initial crawled news

3.2 Parsing:

For parsing relevant data from crawled news, we wrote python code where we first removed the image tags and only kept Bengali words of the news.

মাওয়া ঘাটে খেতে গিয়ে ফেরা হলো না তাঁর
মুন্সিগঞ্জের লৌহজং উপজেলায় প্রাইভেটকার ও মালবাহী ট্রাকের মুখোমুখি সংঘর্ষে এক ব্যক্তি নিহত হয়েছেন আজ সোমবার ভোর পৌনে পাঁচটার দিকে উপজেলার কুমারভোগ এলাকায় ঢাকা মাওয়া মহাসড়কে এ দুর্ঘটনা ঘটে নিহত ব্যক্তির নাম ফুয়াদ হোসেন ৪০ তাঁর বাসা রাজধানীর উত্তরায় যুক্তরাষ্ট্রের নাগরিকত্ব ছিল তাঁর এ ঘটনায় চারজন আহত হয়েছেন আহত ব্যক্তিদের ঢাকা মেডিকেল কলেজ হাসপাতালে পাঠানো হয়েছে প্রথম আলোর হিসাবে এই নিয়ে ৬৫৫ দিনে সড়কে ঝরল ৫ হাজার ৭৩২ জনের প্রাণ লৌহজং থানার পরিদর্শক তদন্ত মো রাজিব খান প্রথম আলোকে জানান ঢাকা থেকে রাতে খাওয়ার উদ্দেশে ফুয়াদসহ পাঁচজন প্রাইভেটকারে করে মাওয়া ঘাটে আসেন ভোরে রাজধানীর উদ্দেশে ফিরে যাচ্ছিলেন তাঁরা পৌনে পাঁচটার দিকে কুমারভোগ পুরোনো বাসস্ট্যান্ড এলাকায় পৌঁছালে মাওয়া ঘাটগামী মালবাহী একটি ট্রাকের সঙ্গে তাঁদের প্রাইভেটকারটির মুখোমুখি সংঘর্ষ হয় এতে ঘটনাস্থলেই ফুয়াদ মারা যান প্রাইভেটকারে থাকা অন্যদের উদ্ধার করে উপজেলা স্বাস্থ্য কমপ্লেক্সে নিয়ে যাওয়া হয় পরে অবস্থা গুরুতর হওয়ায় তাঁদের ঢাকা মেডিকেল কলেজ হাসপাতালে পাঠানো হয়
১০ ডিসেম্বর ২০১৮, ১১:১১

সিলিন্ডার বিস্ফোরণে একই পরিবারের ৩ জন দগ্ধ
ঢাকার অদূরে আশুলিয়ার বাইপাইলে একটি বাসায় গ্যাস সিলিন্ডার বিস্ফোরণে স্বামী স্ত্রী ও ছেলে দগ্ধ হয়েছেন গুরুতর অবস্থায় তাঁদের ঢাকা মেডিকেল কলেজ হাসপাতালের বার্ন ইউনিটে ভর্তি করা হয়েছে আজ শনিবার ভোর পাঁচটার দিকে এ ঘটনা ঘটে দগ্ধ তিনজন হলেন আকরাম ৩৫ তাঁর স্ত্রী লাভলী ২৫ ও ছেলে হোসেন মোহাম্মদ আশিক ৮ ঘটনার সত্যতা নিশ্চিত করেছেন ঢাকা মেডিকেল কলেজ হাসপাতালের পুলিশ ফাঁড়ির ইনচার্জ বাচ্চু মিয়া তিনি জানান বাইপাইলে এসএ পরিবহনের কার্যালয়ের পাশে মোস্তফা নামের এক ব্যক্তির বাসায় ভাড়া থাকত পরিবারটি আজ ভোর পাঁচটার দিকে চুলা জ্বালানোর সময় সিলিন্ডার বিস্ফোরণ ঘটলে আকরাম ও তাঁর স্ত্রী সন্তান দগ্ধ হন চিকিৎসকের বরাত দিয়ে বাচ্চু মিয়া জানান আকরামের ৯০ শতাংশ স্ত্রীর ৬০ শতাংশ এবং ছেলে আশিকের ৫ শতাংশ পুড়ে গেছে
০১ ডিসেম্বর ২০১৮, ০৯:১০

Figure 7: Parsed news

3.3 ASCII to Unicode Conversion:

After analysis, we found out that news on the Prothom-Alo website was not consistent. Some of the news was written in the "ASCII" format and some were written in "UNICODE".

Word	ASCII Spelling	Unicode Spelling
মৌলভীবাজার	ম, ে, ৗ, ল, ভ, ী, ব, া, জ, া, র	ম, ৌ, ল, ভ, ী, ব, া, জ, া, র
পিরোজপুর	প, ি, র, ে, া, জ, প, ু, র	প, ি, র, ো, জ, প, ু, র
নারায়ণগঞ্জ	ন, া, র, া, য, ়, ণ, গ, ঞ, ্, জ	ন, া, র, া, য়, ণ, গ, ঞ, ্, জ
ব্রাহ্মণবাড়িয়া	ব, ্, র, া, হ, ্, ম, ণ, ব, া, ড, ়, ি, য, ়, া	ব, ্, র, া, হ, ্, ম, ণ, ব, া, ড়, ি, য়, া
নোয়াখালী	ন, ে, া, য, ়, া, খ, া, ল, ী	ন, ো, য়, া, খ, া, ল, ী
কুষ্টিয়া	ক, ু, ষ, ্, ট, ি, য, ়, া	ক, ু, ষ, ্, ট, ি, য়, া
বগুড়া	ব, গ, ু, ড, ়, া	ব, গ, ু, ড়, া

Figure 8: ASCII and UNICODE font spelling of some words

We can see some examples where the same words were differently spelled (Figure-8). So, in order to fix the problem, we converted the whole data into the Unicode programmatically.

3.4 Eliminating stop words and Indexing to the root word:

The Bengali language has about 1,00,000 unique words and we can categorize them into 4 groups. First one is the Totsomo(তৎসম) which is directly re-borrowed from the Sanskrit. Almost 50,000 words are considered to be Totsomo. 21,000 are Todbhob(তদ্ভব) which means native words. Rest words are either Bideshi(বিদেশী) which are foreign borrowed or Deshi(দেশী) which are Austroasiatic borrowed[4]. All Bengali words are also categorized into five parts of speech, they are: Noun, Pronoun, Adjective, Conjunction and Verb. We use words to make a meaningful sentence which can express any thought, feelings etc. Among the words which make a complete sentence, some of them provide us information and the rest of them to help those meaningful words to express information appropriately such as auxiliary verbs and conjunctions. We refer them as Stop words. For better understanding, we can see the following example.

সাভারের আমিনবাজারে ট্রাকচাপায় এক ছাত্র নিহত হয়েছেন। আজ মঙ্গলবার সকাল ১০টার দিকে আমিনবাজারের অনলাইন সিএনজি পাম্পের সামনে, ঢাকা আরিচা মহাসড়কে এ ঘটনা ঘটে। নিহত ছাত্রের নাম সোহেল শেখ(২২), তিনি ঢাকার কবি নজরুল সরকারি কলেজের প্রাণিবিদ্যা বিভাগের দ্বিতীয় বর্ষের ছাত্র ছিলেন। তিনি মাদারীপুরের শিবচর উপজেলার গুপ্তেরকান্দির দত্তপাড়ার আনসার শেখের ছেলে। ঢাকার কামরাঙ্গীরচরে রসুলপুর এলাকায় থাকতেন। তিনি এই নিয়ে প্রথম আলোর হিসাবে ৬২১ দিনে দেশে সড়ক দুর্ঘটনায় নিহত হয়েছেন ৫ হাজার ৪৯৩ জন। পুলিশ ও পরিবার সূত্রে জানা গেছে সকালে ঢাকা থেকে সাভারে মোটরসাইকেলে করে যাচ্ছিলেন সোহেল শেখ, আমিনবাজারের অনলাইন সিএনজি পাম্পের সামনে পৌঁছালে একটি ট্রাক তাঁর মোটরসাইকেলটিকে চাপা দেয় এতে ঘটনাস্থলেই তাঁর মৃত্যু হয়। সাভার থানার উপপরিদর্শক এসআই অখিল রঞ্জন সরকার এ তথ্যের সত্যতা নিশ্চিত করেছেন। তিনি জানান লাশ থানায় নিয়ে যাওয়া হয়েছে নিহত ব্যক্তির ছোট ভাই জুয়েল শেখ ময়নাতদন্ত ছাড়াই লাশ গ্রহণের আবেদন করেছেন।

Here we marked some of the stop/ignorable words. They are listed below:

{

এক, হয়েছেন, দিক, এ, ঘটে, তিনি, ছিলেন, এলাকায়, থাকতেন, এই, নিয়ে, জন,জানা, গেছে, থেকে, করে, যাচ্ছিলেন, পৌঁছালে, একটি, তাঁর, দেয়, এতে, করেছেন, জানান, যাওয়া, হয়েছে, ছাড়াই

}

And the rest words are considered as Keyword. We have studied more than 2000 news, listed more than 1500 stop words and more than 5000 root words in a text file.

But storing root words was not as easy as stop words. There were some problems. For example “যাচ্ছিলেন” and “যাচ্ছে” ; “গেছে” , “গিয়েছে” and “গিয়েছেন” ; “করেছেন” , “করছে” and “করে”; “থাকতেন” and “থাকে” ; “হয়েছেন” and “হয়েছে”; They are in different form but their meaning are same. To solve these we read about Bengali grammatical rules and they are পদ, প্রকৃতি, প্রত্যয়, উপসর্গ, অনুসর্গ, কারক, বিভক্তি ও সন্ধি-বিচ্ছেদ.

Table 1: Root and Stop words

STOP WORDS	ROOT WORDS
অতএব	চিকিৎসকসহ - চিকিৎসক
সংগৃহীত	চালককে - চালক
হয়	থানায় - থানা
গেছে	কোম্পানিটিতে - কোম্পানি
বলা	গ্রেফতারের - গ্রেফতার
জানান	বাসটিকে - বাস
বলেন	অনুরোধে - অনুরোধ
যায়	শিক্ষার্থীর - শিক্ষার্থী
অথবা	আগেই - আগে
অনেক	হারায় - হারানো
আমরা	যুক্তরাষ্ট্রের - যুক্তরাষ্ট্র
আমি	উত্তরপূর্বে - উত্তরপূর্ব
উচিত	বিশিষ্টজনদের - বিশিষ্টজন
ইহা	স্বভাবিকভাবেই - স্বভাবিক
এবং	পত্রপত্রিকায় - পত্রপত্রিকা
এত	বিরোধীদলীয় - বিরোধীদল

Using stop words and root words files, we removed stop words from our modified parsed news and indexed root words. Data that we got after processing was used as our final data.

মাওয়া ঘাটে ফেরা
মুন্সিগঞ্জের লৌহজং উপজেলা প্রাইভেটকার মালবাহী ট্রাক সংঘর্ষ ব্যক্তি নিহত ভোর পাঁচটার কুমারভোগ ঢাকা মাওয়া মহাসড়কে দুর্ঘটনা নিহত ব্যক্তি ফুয়াদ ৪০ বাসা রাজধানী উত্তরায় যুক্তরাষ্ট্র নাগরিকত্ব ঘটনা চারজন আহত আহত ঢাকা আলোর ৬৫৫ সড়ক ঝরল ৫ ৭৩২ প্রাণ লৌহজং পরিদর্শক তদন্ত রাজিব ঢাকা খাওয়া উদ্দেশে ফুয়াদসহ পাঁচজন প্রাইভেটকারে মাওয়া ঘাটে রাজধানী উদ্দেশে পাঁচটার কুমারভোগ পুরোনো বাসস্ট্যান্ড মাওয়া ঘাটগামী মালবাহী ট্রাক প্রাইভেটকারটির সংঘর্ষ ঘটনাস্থল ফুয়াদ মারা প্রাইভেটকারে অন্যদের উপজেলা কমপ্লেক্স গুরুতর ঢাকা
১০ ডিসেম্বর ২০১৮, ১১:১১

সিলিন্ডার বিস্ফোরণে পরিবার ৩ দগ্ধ
ঢাকা অদূরে আশুলিয়ার বাইপাইলে বাসা গ্যাস সিলিন্ডার বিস্ফোরণে স্বামী স্ত্রী দগ্ধ গুরুতর ঢাকা বার্ন ইউনিট ভোর পাঁচটার ঘটনা দগ্ধ তিনজন আকরাম ৩৫ স্ত্রী লাভলী ২৫ মোহাম্মদ আশিক ৮ ঘটনা ঢাকা পুলিশ ফাঁড়ির ইনচার্জ বাচ্চু বাইপাইলে পরিবহন কার্যালয় মোস্তফা ব্যক্তি বাসা থাকত পরিবার ভোর পাঁচটার চুলা জ্বালানোর সিলিন্ডার বিস্ফোরণ ঘটলে আকরাম স্ত্রী সন্তান দগ্ধ চিকিৎসকের বরাত বাচ্চু আকরামের ৯০ শতাংশ স্ত্রী ৬০ শতাংশ আশিকের ৫ শতাংশ পুড়ে
০১ ডিসেম্বর ২০১৮, ০৯:১০

তেজগাঁওয়ে মালবাহী ট্রেন বগি লাইনচ্যুত
রাজধানী তেজগাঁও রেলগেট মালবাহী ট্রেন বগি লাইনচ্যুত দুর্ঘটনা তেজগাঁও রেলগেটের গেটম্যান রেজাউল মালবাহী ট্রেনটি রাজধানী কমলাপুর চট্টগ্রাম ভোররাত চারটার ট্রেনটির ইঞ্জিনের বগি লাইনচ্যুত বগির চাকা রেললাইনের বগি লাইনচ্যুত সীমিত আকারে ট্রেন চলাচল রেজাউল ঢাকা রেলথানার কর্মকর্তা ইয়াসিন ফারুক ১০টার বগি লাইনচ্যুত ইতিমধ্যে বগি বগি তোলা
১৪ নভেম্বর ২০১৮, ১০:৫৮

Figure 9: Final parsed data

The actual news paragraphs were kept apart from our final data in another file without heading and date for further processing steps. We clustered all the news using this file and our final data.

মুন্সিগঞ্জের লৌহজং উপজেলা প্রাইভেটকার মালবাহী ট্রাক সংঘর্ষ ব্যক্তি নিহত ভোর পাঁচটার কুমারভোগ ঢাকা মাওয়া মহাসড়কে দুর্ঘটনা নিহত ব্যক্তি ফুয়াদ ৪০ বাসা রাজধানী উত্তরায় যুক্তরাষ্ট্র নাগরিকত্ব ঘটনা চারজন আহত আহত ঢাকা আলোর ৬৫৫ সড়ক ঝরল ৫ ৭৩২ প্রাণ লৌহজং পরিদর্শক তদন্ত রাজিব ঢাকা খাওয়া উদ্দেশে ফুয়াদসহ পাঁচজন প্রাইভেটকারে মাওয়া ঘাটে রাজধানী উদ্দেশে পাঁচটার কুমারভোগ পুরোনো বাসস্ট্যান্ড মাওয়া ঘাটগামী মালবাহী ট্রাক প্রাইভেটকারটির সংঘর্ষ ঘটনাস্থল ফুয়াদ মারা প্রাইভেটকারে অন্যদের উপজেলা কমপ্লেক্স গুরুতর ঢাকা

ঢাকা অদূরে আশুলিয়ার বাইপাইলে বাসা গ্যাস সিলিন্ডার বিস্ফোরণে স্বামী স্ত্রী দগ্ধ গুরুতর ঢাকা বার্ন ইউনিট ভোর পাঁচটার ঘটনা দগ্ধ তিনজন আকরাম ৩৫ স্ত্রী লাভলী ২৫ মোহাম্মদ আশিক ৮ ঘটনা ঢাকা পুলিশ ফাঁড়ির ইনচার্জ বাচ্চু বাইপাইলে পরিবহন কার্যালয় মোস্তফা ব্যক্তি বাসা থাকত পরিবার ভোর পাঁচটার চুলা জ্বালানোর সিলিন্ডার বিস্ফোরণ ঘটলে আকরাম স্ত্রী সন্তান দগ্ধ চিকিৎসকের বরাত বাচ্চু আকরামের ৯০ শতাংশ স্ত্রী ৬০ শতাংশ আশিকের ৫ শতাংশ পুড়ে

রাজধানী তেজগাঁও রেলগেট মালবাহী ট্রেন বগি লাইনচ্যুত দুর্ঘটনা তেজগাঁও রেলগেটের গেটম্যান রেজাউল মালবাহী ট্রেনটি রাজধানী কমলাপুর চট্টগ্রাম ভোররাত চারটার ট্রেনটির ইঞ্জিনের বগি লাইনচ্যুত বগির চাকা রেললাইনের বগি লাইনচ্যুত সীমিত আকারে ট্রেন চলাচল রেজাউল ঢাকা রেলথানার কর্মকর্তা ইয়াসিন ফারুক ১০টার বগি লাইনচ্যুত

Figure 10: News without heading and date

3.5 Categorizing News

We used three algorithms to categorize our news:

1. Naive Bayes
2. Cosine Similarity
3. Support Vector Machine

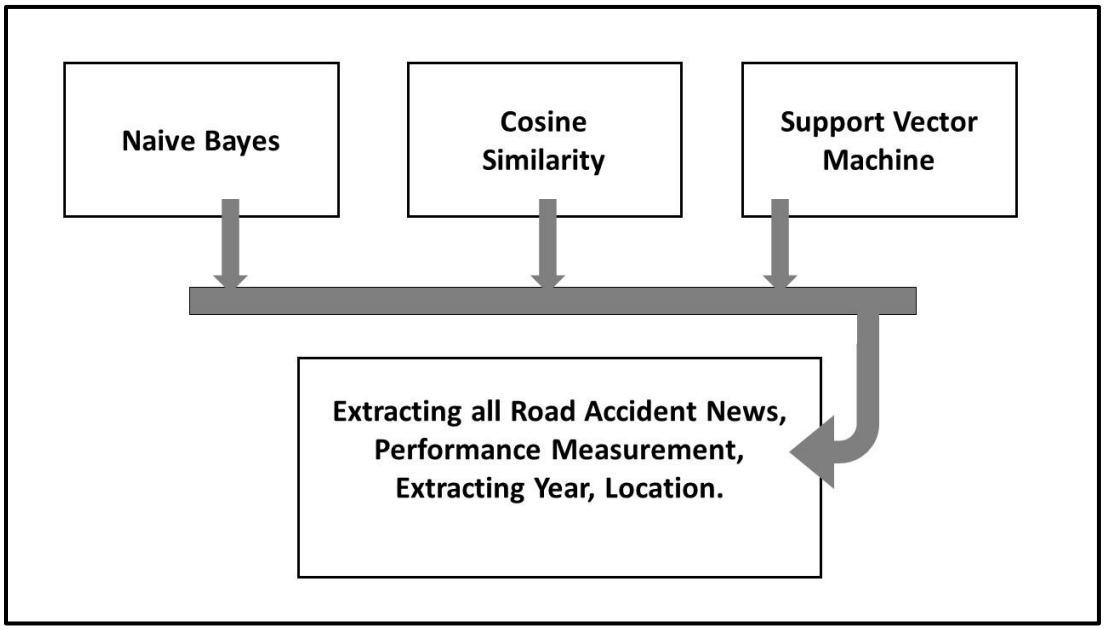


Figure 11: Categorizing news and further procedure

3.6 Classification using Naive Bayes

Firstly, we used Naive Bayes classifier to categorize our modified parsed news. For Naive Bayes train data, we made a file containing 500 news of several types of accidents. The train data was classified as positive and negative. If the news was a road accident, we identified it as "positive" and identified "negative" otherwise. In the negative class, there was some news about accidents also, but these weren't road accidents. Some of them from -

- ভবন ধ্বস (Building collapse)
- বজ্রপাতে মৃত্যু (death from lightning strike)
- জাহাজ/ লঞ্চ/ নৌকাডুবি (Sinking boat/ launch/ ship)
- ট্রেন দুর্ঘটনা (Train Accidents)
- পুকুরে ডুবে মৃত্যু (Drowning in the pond)
- আগুনে পুড়ে মৃত্যু (Burnt to death in fire accidents)
- পশুর আক্রমণে মৃত্যু (Death by animal attack)
- সিলিন্ডার বিস্ফোরণ (cylinder blast)

রংপুর কাউনিয়া হলদিবাড়ি বাস ধাক্কা মোটরসাইকেলের যাত্রী নিহত রংপুর কুড়িগ্রাম মহাসড়কের রেলগেটের আটটার দুর্ঘটনা কাউনিয়া পরিদর্শক এসআই আজিজ ঘটনা নিহত ব্যাক্তি রংপুর গঙ্গাচড়া চরমরনিয়া আবুল কালাম ২২ সোলাইমান ২০ জামালপুর চর শিশুয়া ইসলামপুর আবদুস সাত্তার প্রামাণিক ৭২ প্রত্যক্ষদর্শী ব্যাক্তি ট্রেন মোটরসাইকেল স্টেশন তিনজন আবদুস সাত্তার প্রামাণিক জামালপুর চাওয়া সম্পর্ক আবুল কালাম সোলাইমানের দাদা মোটরসাইকেল চাচাতো রংপুর কাউনিয়া স্টেশন রংপুরে যাত্রী ঢাকা ফিরতে খালি বাস চাপা
positive
গাজীপুর বাইমাইল বিলে নৌকাডুবিতে শিশু মারা বিকেল চারটার গাজীপুর বাইমাইল বিলের নৌকাডুবির ঘটনাটি নিহত চারজনের দুজন সহোদর জয়দেবপুর ভোগড়া পুলিশ এএসআই রেজাউল হক ঘটনা নৌকাডুবিতে নিহত শিশুরা জামালপুর সরিষাবাড়ী সাতরা খোকনের পারভিন ১২ টাঙ্গাইল কাঠতলা সাইফুল সাদিয়া ১০ পাবনার বানপুর নূরনগর মজিবরের সোহাগ ৮ সোহাগের মীম ১০ গাজীপুর করপোরেশনের হরিণাচালা থাকত হরিণাচালা শিশু ঈদে নৌকাভ্রমণের পরিকল্পনা তিনটার পার্শ্ববর্তী বিলে ঘোরার নৌকায় একপর্যায় পানি নিচে কংক্রিটের পিলারে নৌকাটির ধাক্কা লাগে নৌকার ফাটল তলিয়ে সাঁতার শিশু তলিয়ে পাঁচজন বাঁশের খুঁটি চিৎকার থকলে
negative
জয়পুরহাটের উকিলের মোড় ভটভটির ধাক্কা দুজন নিহত ভোর চারটার ঘটনা আহত সাতজন হতাহতরা গরু ব্যবসায়ী নিহতেরা কাওসার ৪৫ মানিক নিহত একজনের দিনাজপুর আরেকজনের চাঁপাইনবাবগঞ্জ ব্যক্তি পুলিশ ভটভটিতে নয়জন গরু ব্যবসায়ী রাজশাহী গরু দিনাজপুরের হিলি হাকিমপুর জয়পুরহাট ধামুইরহাটে সীমান্ত উকিলের মোড়ের ভটভটি নিয়ন্ত্রণ ধাক্কা খায় ঘটনাস্থল কাওসার ৪৫ মারা আহত জয়পুরহাট হাসপাতাল মানিক ২৫ মারা আহত জয়পুরহাট হাসপাতাল চিকিৎসা জয়পুরহাট পরিদর্শক এসআই তদন্ত মুমিনুল হক কোকিলের মোড় দুর্ঘটনা পরপর হাসপাতাল ঘটনা দুজন মারা
positive

Figure 12: Train data of Naive Bayes

Then using our prepared train data, we categorized our news. The probability of a document being in class C_i which we call Posterior Probability of C_i is computed as[18]:

$$\boldsymbol{Posterior\ Probability}\ (\boldsymbol{C_i}) = \boldsymbol{Prior\ Probability}(\boldsymbol{C_i}) * \boldsymbol{Conditional\ Probability}(\frac{\boldsymbol{Word_i}}{\boldsymbol{C_i}})$$

Here we marked some of the stop/ignorable words. They are listed below:

Where,

$$\boldsymbol{Prior\ Probability}(\boldsymbol{C_i}) = \frac{\boldsymbol{N_c}}{\boldsymbol{N_d}}$$

Here,

$$N_c = Total\ Number\ of\ class$$

$$N_d = Total\ documents$$

And,

$$\boldsymbol{Conditional\ Probability}\ \left(\frac{\boldsymbol{word_i}}{\boldsymbol{C_i}}\right) = \boldsymbol{count}\left(\frac{\boldsymbol{word_i}}{\boldsymbol{C_i}}\right) + \frac{\boldsymbol{1}}{\boldsymbol{count}(\boldsymbol{C_i}) + |\boldsymbol{V}|}$$

Here,

$$count\left(\frac{word_i}{C_i}\right) = Count\ of\ word_i\ in\ class\ C_i$$

$$count(C_i) = \ Total\ words\ of\ the\ documents\ in\ C_i\ class$$

$$|V| = Total\ words\ in\ frequendy\ table$$

A frequency table is a table where, how many times a word of the document appears in C class of the train data is stored.

After calculating Posterior Probability of all classes, we find in which class the document belongs to by taking the class with maximum value. So,

$$\boldsymbol{Class\ of\ document = max(Posterior\ Probability(C_1, C_2 \ldots\ldots C_{N_c}))}$$

After Appling Naive Bayes, it gave us the output where the positive class is referred to road accident news and the negative class is referred to other accident news.

মুন্সিগঞ্জের লৌহজং উপজেলা প্রাইভেটকার মালবাহী ট্রাক সংঘর্ষ ব্যক্তি নিহত ভোর পাঁচটার কুমারভোগ ঢাকা মাওয়া মহাসড়কে দুর্ঘটনা নিহত ব্যক্তি ফুয়াদ ৪০ বাসা রাজধানী উত্তরায় যুক্তরাষ্ট্র নাগরিকত্ব ঘটনা চারজন আহত আহত ঢাকা আলোর ৬৫৫ সড়ক ঝরল ৫ ৭৩২ প্রাণ লৌহজং পরিদর্শক তদন্ত রাজিব ঢাকা খাওয়া উদ্দেশে ফুয়াদসহ পাঁচজন প্রাইভেটকারে মাওয়া ঘাটে রাজধানী উদ্দেশে পাঁচটার কুমারভোগ পুরোনো বাসস্ট্যান্ড মাওয়া ঘাটগামী মালবাহী ট্রাক প্রাইভেটকারটির সংঘর্ষ ঘটনাস্থল ফুয়াদ মারা প্রাইভেটকারে অন্যদের উপজেলা কমপ্লেক্স গুরুতর ঢাকা
Road Accident

ঢাকা অদূরে আশুলিয়ার বাইপাইলে বাসা গ্যাস সিলিন্ডার বিস্ফোরণে স্বামী স্ত্রী দগ্ধ গুরুতর ঢাকা বার্ন ইউনিট ভোর পাঁচটার ঘটনা দগ্ধ তিনজন আকরাম ৩৫ স্ত্রী লাভলী ২৫ মোহাম্মদ আশিক ৮ ঘটনা ঢাকা পুলিশ ফাঁড়ির ইনচার্জ বাচ্চু বাইপাইলে পরিবহন কার্যালয় মোস্তফা ব্যক্তি বাসা থাকত পরিবার ভোর পাঁচটার চুলা জ্বালানোর সিলিন্ডার বিস্ফোরণ ঘটলে আকরাম স্ত্রী সন্তান দগ্ধ চিকিৎসকের বরাত বাচ্চু আকরামের ৯০ শতাংশ স্ত্রী ৬০ শতাংশ আশিকের ৫ শতাংশ পুড়ে
Others

রাজধানী সবুজবাগ ভবনে বিদ্যুৎস্পৃষ্ট নিহত শুভ চন্দ্র ২৬ মোহাম্মদ শুভ ২৮ ৯টার ঘটনা সবুজবাগ এসআই সাইফুর রহমান নিহত দোকানি সোহরাব আটটার সবুজবাগ তৃতীয় তলার ব্যালকনিতে গ্রিল লাগানো বিদ্যুৎস্পৃষ্ট দুজন গুরুতর আহত মুগদা চিকিৎসক মৃত ঘোষণা নিহত শুভ চন্দ্র পরিবারসহ মুগদার ব্যাংক কলোনিতে নীলফামারী নিহত মোহাম্মদ শুভ সবুজবাগের কাঠেরপুলে মেসে ঢাকা দোহার উপজেলা শ্রমিক
Others

Figure 13: Output of Naive Bayes

We also created another output file where we kept all the news of only road accident along its heading and date.

মাওয়া ঘাটে ফেরা
মুন্সিগঞ্জের লৌহজং উপজেলা প্রাইভেটকার মালবাহী ট্রাক সংঘর্ষ ব্যক্তি নিহত ভোর পাঁচটার কুমারভোগ ঢাকা মাওয়া মহাসড়কে দুর্ঘটনা নিহত ব্যক্তি ফুয়াদ ৪০ বাসা রাজধানী উত্তরায় যুক্তরাষ্ট্র নাগরিকত্ব ঘটনা চারজন আহত আহত ঢাকা আলোর ৬৫৫ সড়ক ঝরল ৫ ৭৩২ প্রাণ লৌহজং পরিদর্শক তদন্ত রাজিব ঢাকা খাওয়া উদ্দেশে ফুয়াদসহ পাঁচজন প্রাইভেটকারে মাওয়া ঘাটে রাজধানী উদ্দেশে পাঁচটার কুমারভোগ পুরোনো বাসস্ট্যান্ড মাওয়া ঘাটগামী মালবাহী ট্রাক প্রাইভেটকারটির সংঘর্ষ ঘটনাস্থল ফুয়াদ মারা প্রাইভেটকারে অন্যদের উপজেলা কমপ্লেক্স গুরুতর ঢাকা
১০ ডিসেম্বর ২০১৮, ১১:১১

নরসিংদীতে বাস সংঘর্ষ নিহত ৩
নরসিংদী শিবপুরে বাস সংঘর্ষ চালকসহ তিনজন নিহত আহত ১১টার শিবপুরের সৈয়দনগর বাসস্ট্যান্ডে ঢাকা সিলেট মহাসড়কে দুর্ঘটনা নিহত নরসিংদী বাসচালক আনোয়ার ৪৫ শিবপুরের ইটাখোলার আলাউদ্দিন ৬০ রেহেনা ৪৫ প্রত্যক্ষদর্শী ভাষ্য ঢাকা ভৈরব রয়েল পরিবহন যাত্রীবাহী বাস শিবপুরের সৈয়দনগর বাসস্ট্যান্ড সড়ক দাঁড়িয়েছিল যাত্রীবাহী লোকাল বাস রয়েল পরিবহন যাত্রীবাহী বাস চলন্ত বাসকে পাশ দাঁড়ানো যাত্রীবাহী লোকাল বাস রয়েল পরিবহন বাসটির সংঘর্ষ লোকাল বাস চালকসহ দুজন ঘটনাস্থল নিহত লোকজন আহত নরসিংদী মারা নরসিংদী আবাসিক চিকিৎসা কর্মকর্তা এন এম মিজানুর রহমান সড়ক দুর্ঘটনা আহত ২১
০৭ ডিসেম্বর ২০১৮, ১৫:৪৭

Figure 14: Output of positive class from Naive Bayes

3.6.1 The Accuracy of Naive Bayes

We read 500 news manually to measure the accuracy. We listed the outcomes of classes as follows:

- True Positive (TP): 248
- True Negative (TN): 238
- False positive (FP): 13
- False Negative (FN): 0

Here,

TP: News of road accident classified as road accident

TN: News of other accident classified as other accident

FP: News of other accident classified as road accident

FN: News of road accident classified as other accident

Using the equation for measuring accuracy, we found the accuracy of Naive Bayes is 97.2%

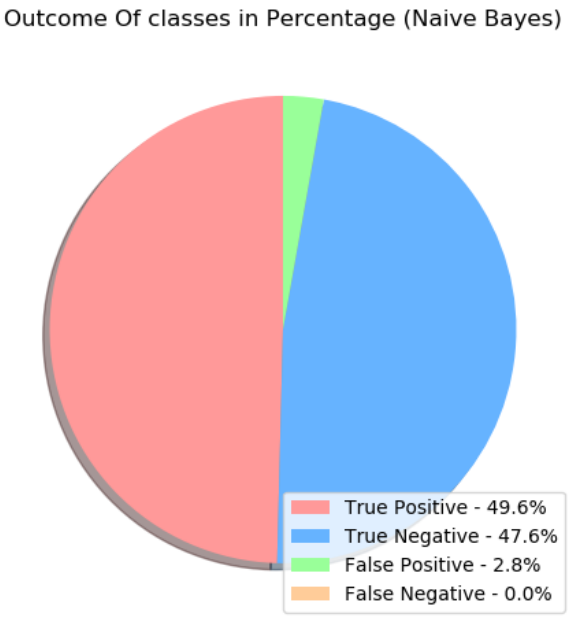


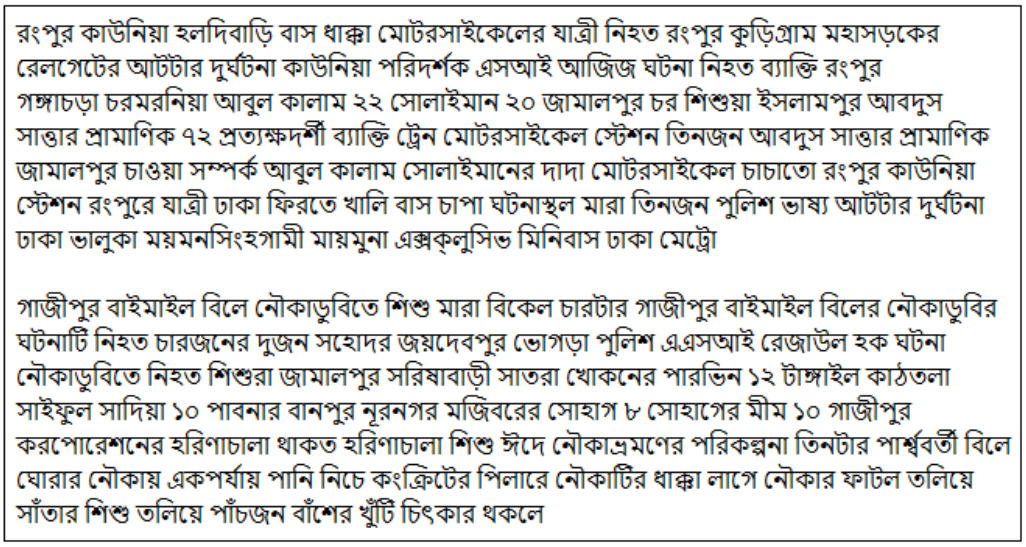
Figure 15: Pie-chart of Naive Bayes's outcomes (Classes)

3.7 Classification using Cosine-Similarity

To implement Cosine-Similarity, we used a python library called scikit-learn. We used cosine_similarity, TfidfVectorizer, and TfidfTransformer from scikit-learn to implement the cosine similarity code.

To create train data, we took 500 news of different types of accidents; inside the implementation code, we wrote the index numbers of the road accident news. And there were several types of accident news in the Cosine-Similarity train data. Some of them were:

- ভবন ধ্বস (Building collapse)
- বজ্রপাতে মৃত্যু (death from lightning strike)
- জাহাজ/ লঞ্চ/ নৌকাডুবি (Sinking boat/ launch/ ship)
- ট্রেন দুর্ঘটনা (Train Accidents)
- পুকুরে ডুবে মৃত্যু (Drowning in the pond)
- আগুনে পুড়ে মৃত্যু (Burnt to death in fire accident)
- পশুর আক্রমণে মৃত্যু (Death by animal attack)
- সিলিন্ডার বিস্ফোরণ (cylinder blast)



রংপুর কাউনিয়া হলদিবাড়ি বাস ধাক্কা মোটরসাইকেলের যাত্রী নিহত রংপুর কুড়িগ্রাম মহাসড়কের রেলগেটের আটটার দুর্ঘটনা কাউনিয়া পরিদর্শক এসআই আজিজ ঘটনা নিহত ব্যাক্তি রংপুর গঙ্গাচড়া চরমরনিয়া আবুল কালাম ২২ সোলাইমান ২০ জামালপুর চর শিশুয়া ইসলামপুর আবদুস সাত্তার প্রামাণিক ৭২ প্রত্যক্ষদর্শী ব্যাক্তি ট্রেন মোটরসাইকেল স্টেশন তিনজন আবদুস সাত্তার প্রামাণিক জামালপুর চাওয়া সম্পর্ক আবুল কালাম সোলাইমানের দাদা মোটরসাইকেল চাচাতো রংপুর কাউনিয়া স্টেশন রংপুরে যাত্রী ঢাকা ফিরতে খালি বাস চাপা ঘটনাস্থল মারা তিনজন পুলিশ ভাষ্য আটটার দুর্ঘটনা ঢাকা ভালুকা ময়মনসিংহগামী মায়মুনা এক্সক্লুসিভ মিনিবাস ঢাকা মেট্রো

গাজীপুর বাইমাইল বিলে নৌকাডুবিতে শিশু মারা বিকেল চারটার গাজীপুর বাইমাইল বিলের নৌকাডুবির ঘটনাটি নিহত চারজনের দুজন সহোদর জয়দেবপুর ভোগড়া পুলিশ এএসআই রেজাউল হক ঘটনা নৌকাডুবিতে নিহত শিশুরা জামালপুর সরিষাবাড়ী সাতরা খোকনের পারভিন ১২ টাঙ্গাইল কাঠতলা সাইফুল সাদিয়া ১০ পাবনার বানপুর নূরনগর মজিবরের সোহাগ ৮ সোহাগের মীম ১০ গাজীপুর করপোরেশনের হরিণাচালা থাকত হরিণাচালা শিশু ঈদে নৌকাভ্রমণের পরিকল্পনা তিনটার পার্শ্ববর্তী বিলে ঘোরার নৌকায় একপর্যায় পানি নিচে কংক্রিটের পিলারে নৌকাটির ধাক্কা লাগে নৌকার ফাটল তলিয়ে সাঁতার শিশু তলিয়ে পাঁচজন বাঁশের খুঁটি চিৎকার থকলে

Figure 16: train data of Cosine-Similarity

The first news (Figure-16) was indexed as the positive and the second one was negative.

Then, by the TfidfVectorizer, TfidfTransformer, we have transformed the data into vector form and then took all the test news. The scikit-learn took all the words from test news and train news as features. The news was matched one by one. We checked for each news N, to which train data it is matching becomes the maximum. And finally, we checked if the maximum matched index is in our listed road accident index or not.

After completing the cosine-similarity, we created two output files. One consisting of each of the test news and output as road accident or others.

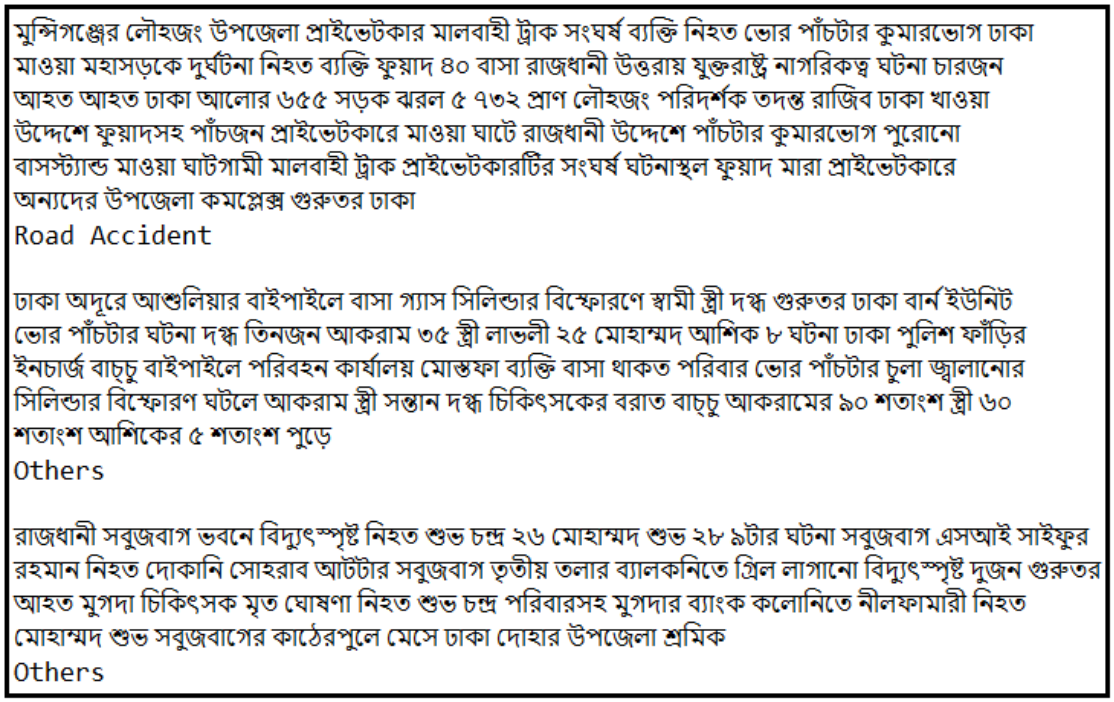


Figure 17: Output of Cosine-Similarity

And another file consisting of only road accident news.

মুন্সিগঞ্জের লৌহজং উপজেলা প্রাইভেটকার মালবাহী ট্রাক সংঘর্ষ ব্যক্তি নিহত ভোর পাঁচটার কুমারভোগ ঢাকা মাওয়া মহাসড়কে দুর্ঘটনা নিহত ব্যক্তি ফুয়াদ ৪০ বাসা রাজধানী উত্তরায় যুক্তরাষ্ট্র নাগরিকত্ব ঘটনা চারজন আহত আহত ঢাকা আলোর ৬৫৫ সড়ক ঝরল ৫ ৭৩২ প্রাণ লৌহজং পরিদর্শক তদন্ত রাজিব ঢাকা খাওয়া উদ্দেশে ফুয়াদসহ পাঁচজন প্রাইভেটকারে মাওয়া ঘাটে রাজধানী উদ্দেশে পাঁচটার কুমারভোগ পুরোনো বাসস্ট্যান্ড মাওয়া ঘাটগামী মালবাহী ট্রাক প্রাইভেটকারটির সংঘর্ষ ঘটনাস্থল ফুয়াদ মারা প্রাইভেটকারে অন্যদের উপজেলা কমপ্লেক্স গুরুতর ঢাকা

রাজশাহীর পবা উপজেলা যাত্রীবাহী বাস ধাক্কা লাগার ঘটনা দুজন নিহত আহত চারজন চালক নিয়ন্ত্রণ হারালে বাস ধাক্কা খায় রাত তিনটার পবা হরিপুরে দুর্ঘটনা নিহত শওকত আলী ২৬ ৪৫ শওকত আলীর চাঁপাইনবাবগঞ্জে বাস যাত্রী বাস সুপারভাইজারের দায়িত্ব রাজশাহী নগরের রাজপাড়া কেশবপুর আহত চারজনই বাস যাত্রী রাজশাহী অবস্থার উন্নতি দুজনকে ছাড়পত্র চিকিৎসক রাজশাহী নগরের দামকুড়া কর্মকর্তা লতিফ দেশ ট্রাভেলসের বাস ঢাকা চাঁপাইনবাবগঞ্জ রাজশাহীতে যাত্রী চাঁপাইনবাবগঞ্জের যাত্রী রাত তিনটার পবা হরিপুর বাস চালক নিয়ন্ত্রণ হারালে বাস ধাক্কা খায় বাস সুপারভাইজার যাত্রী শওকত আলী

ফেনীতে কাভার্ড ভ্যান ধাক্কা মোটরসাইকেল নারী নিহত স্বামী মেয়েসহ চারজন আহত ১০টার ফেনী বিসিক শিল্পনগরীর চাড়িপুর ঢাকা চট্টগ্রাম মহাসড়কে দুর্ঘটনা নারী রোকসানা পুলিশ লোকজন কুমিল্লার চৌদ্দগ্রাম আলকরা শিলুরী মোটরসাইকেল চড়ে স্ত্রী সন্তান ফেনী বাসা ফিরছিলেন মাদ্রাসাশিক্ষক আবুল হাসেম বিসিক শিল্পনগরীর চাড়িপুর মহাসড়কে বিপরীত দিক দ্রুতগামী কাভার্ড ভ্যান মোটরসাইকেলের ধাক্কা লাগে মোটরসাইকেলসহ আবুল হাসেম ৪০ স্ত্রী রোকসানা ২৮ সন্তান মাহফুজা ১০ মাহমুদা ৮ মাহবুবা ৪ আহত লোকজন গুরুতর আহত চিকিৎসক রোকসানা আক্তারকে মৃত ঘোষণা আহত ফেনীর মহিপাল কর্মকর্তা আউয়াল ঘটনা পুলিশ

Figure 18: Only positive class output of Cosine-Similarity

3.7.1 The Accuracy of cosine similarity

We read 500 news manually to measure the accuracy. We listed the outcomes of classes as follows:

- True Positive (TP): 219
- True Negative (TN): 226
- False positive (FP): 27
- False Negative (FN): 28

Here,

TP: News of road accident classified as road accident

TN: News of other accident classified as other accident

FP: News of other accident classified as road accident

FN: News of road accident classified as other accident

Using the equation for measuring accuracy we found the accuracy of Cosine Similarity is 89%

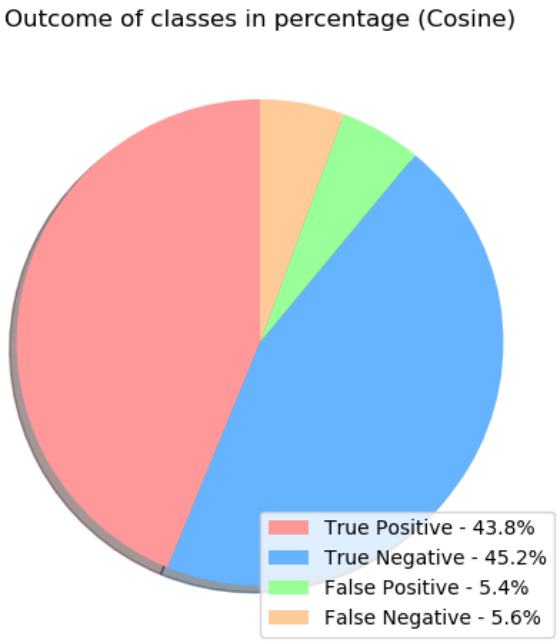


Figure 19: Pie-chart of Cosine Similarity's outcomes (Classes)

3.8 Classification using Support Vector Machine (SVM)

For SVM train data, we also made a file containing 500 news of several types of accident. We classified the train data as positive and negative. If the news was a Road Accident, we identified it as “positive” and identified “negative” otherwise.

In the negative class, there was also some news about accidents, but these were not road accident. Some of these were from-

- ভবন ধ্বস (Building collapse)
- বজ্রপাতে মৃত্যু (died of a lightning strike)
- জাহাজ/ লঞ্চ/ নৌকাডুবি (Sinking boat/ launch/ ship)
- ট্রেন দুর্ঘটনা (Train Accident)
- পুকুরে ডুবে মৃত্যু (Drowning in the pond)
- আগুনে পুড়ে মৃত্যু (Burnt to death)
- পশুর আক্রমণে মৃত্যু (Death by animal attack)
- সিলিন্ডার বিস্ফোরণ (cylinder blast)

রংপুর কাউনিয়া হলদিবাড়ি বাস ধাক্কা মোটরসাইকেলের যাত্রী নিহত রংপুর কুড়িগ্রাম মহাসড়কের রেলগেটের আটটার দুর্ঘটনা কাউনিয়া পরিদর্শক এসআই আজিজ ঘটনা নিহত ব্যাক্তি রংপুর গঙ্গাচড়া চরমরনিয়া আবুল কালাম ২২ সোলাইমান ২০ জামালপুর চর শিশুয়া ইসলামপুর আবদুস সাত্তার প্রামাণিক ৭২ প্রত্যক্ষদর্শী ব্যাক্তি ট্রেন মোটরসাইকেল স্টেশন তিনজন আবদুস সাত্তার প্রামাণিক জামালপুর চাওয়া সম্পর্ক আবুল কালাম সোলাইমানের দাদা মোটরসাইকেল চাচাতো রংপুর কাউনিয়া স্টেশন রংপুরে যাত্রী ঢাকা ফিরতে খালি বাস চাপা
positive
গাজীপুর বাইমাইল বিলে নৌকাডুবিতে শিশু মারা বিকেল চারটার গাজীপুর বাইমাইল বিলের নৌকাডুবির ঘটনাটি নিহত চারজনের দুজন সহোদর জয়দেবপুর ভোগড়া পুলিশ এএসআই রেজাউল হক ঘটনা নৌকাডুবিতে নিহত শিশুরা জামালপুর সরিষাবাড়ী সাতরা খোকনের পারভিন ১২ টাঙ্গাইল কাঠতলা সাইফুল সাদিয়া ১০ পাবনার বানপুর নূরনগর মজিবরের সোহাগ ৮ সোহাগের মীম ১০ গাজীপুর করপোরেশনের হরিণাচালা থাকত হরিণাচালা শিশু ঈদে নৌকাভ্রমণের পরিকল্পনা তিনটার পার্শ্ববর্তী বিলে ঘোরার নৌকায় একপর্যায় পানি নিচে কংক্রিটের পিলারে নৌকাটির ধাক্কা লাগে নৌকার ফাটল তলিয়ে সাঁতার শিশু তলিয়ে পাঁচজন বাঁশের খুঁটি চিৎকার থকলে
negative
জয়পুরহাটের উকিলের মোড় ভটভটির ধাক্কা দুজন নিহত ভোর চারটার ঘটনা আহত সাতজন হতাহতরা গরু ব্যবসায়ী নিহতেরা কাওসার ৪৫ মানিক নিহত একজনের দিনাজপুর আরেকজনের চাঁপাইনবাবগঞ্জ ব্যক্তি পুলিশ ভটভটিতে নয়জন গরু ব্যবসায়ী রাজশাহী গরু দিনাজপুরের হিলি হাকিমপুর জয়পুরহাট ধামুইরহাটে সীমান্ত উকিলের মোড়ের ভটভটি নিয়ন্ত্রণ ধাক্কা খায় ঘটনাস্থল কাওসার ৪৫ মারা আহত জয়পুরহাট হাসপাতাল মানিক ২৫ মারা আহত জয়পুরহাট হাসপাতাল চিকিৎসা জয়পুরহাট পরিদর্শক এসআই তদন্ত মুমিনুল হক কোকিলের মোড় দুর্ঘটনা পরপর হাসপাতাল ঘটনা দুজন মারা
positive

Figure 20: SVM Classified Train data

We used scikit-learn tools to test our data using the Support Vector Machines. We needed to create a pipeline for our train data and test data. For this reason, we have created a news class that will contain test data, its target class serial number and the target class name.

There were two target classes: positive and negative.

For implementing SVM, scikit-learn also created vectorization of the input data. Rather than TfidfVectorizer and TfidfTransformer, we also used MultinomialNB here.

We used the scikit-learn linear SGDClassifier to classify our news using SVM.

We created a pipeline for our test data and train data with SGD (Stochastic Gradient Descent) classifier, TfidfVectorizer and TfIdfTransformer.

And from the comparison, we found the predicted class from the SGDClassifier predict function by providing our test data there.

After Appling Support Vector Machines, we found the output where the positive class is referred to- Road Accident and the negative class is referred to- Others.

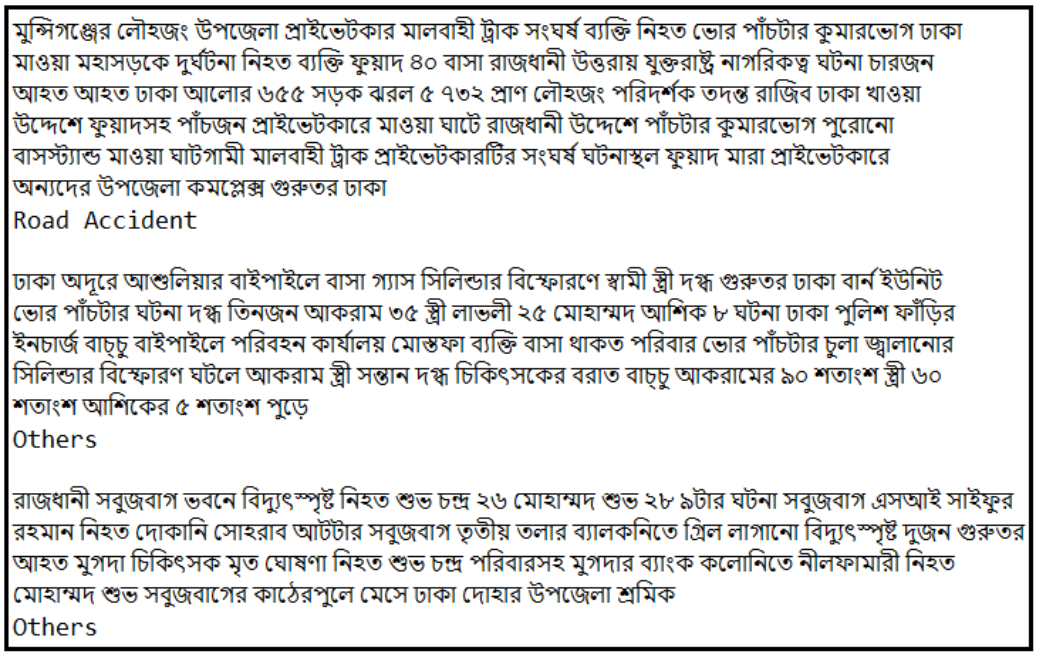
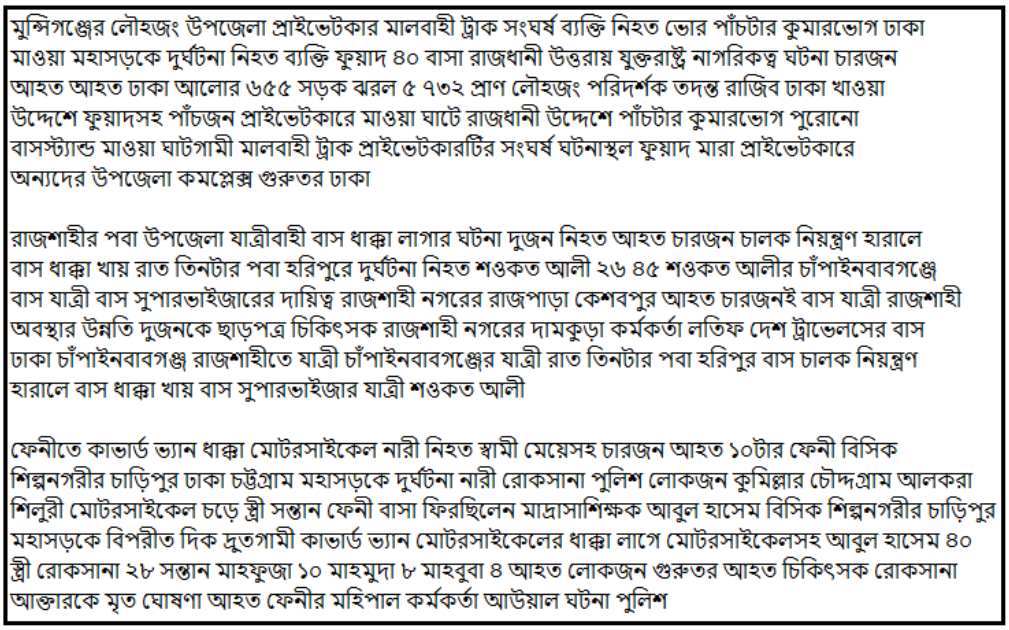


Figure 21: Output of SVM classifier

We also created another output file where we kept all the news which output was road accident


মুন্সিগঞ্জের লৌহজং উপজেলা প্রাইভেটকার মালবাহী ট্রাক সংঘর্ষ ব্যক্তি নিহত ভোর পাঁচটার কুমারভোগ ঢাকা মাওয়া মহাসড়কে দুর্ঘটনা নিহত ব্যক্তি ফুয়াদ ৪০ বাসা রাজধানী উত্তরায় যুক্তরাষ্ট্র নাগরিকত্ব ঘটনা চারজন আহত আহত ঢাকা আলোর ৬৫৫ সড়ক ঝরল ৫ ৭৩২ প্রাণ লৌহজং পরিদর্শক তদন্ত রাজিব ঢাকা খাওয়া উদ্দেশে ফুয়াদসহ পাঁচজন প্রাইভেটকারে মাওয়া ঘাটে রাজধানী উদ্দেশে পাঁচটার কুমারভোগ পুরোনো বাসস্ট্যান্ড মাওয়া ঘাটগামী মালবাহী ট্রাক প্রাইভেটকারটির সংঘর্ষ ঘটনাস্থল ফুয়াদ মারা প্রাইভেটকারে অন্যদের উপজেলা কমপ্লেক্স গুরুতর ঢাকা

রাজশাহীর পবা উপজেলা যাত্রীবাহী বাস ধাক্কা লাগার ঘটনা দুজন নিহত আহত চারজন চালক নিয়ন্ত্রণ হারালে বাস ধাক্কা খায় রাত তিনটার পবা হরিপুরে দুর্ঘটনা নিহত শওকত আলী ২৬ ৪৫ শওকত আলীর চাঁপাইনবাবগঞ্জে বাস যাত্রী বাস সুপারভাইজারের দায়িত্ব রাজশাহী নগরের রাজপাড়া কেশবপুর আহত চারজনই বাস যাত্রী রাজশাহী অবস্থার উন্নতি দুজনকে ছাড়পত্র চিকিৎসক রাজশাহী নগরের দামকুড়া কর্মকর্তা লতিফ দেশ ট্রাভেলসের বাস ঢাকা চাঁপাইনবাবগঞ্জ রাজশাহীতে যাত্রী চাঁপাইনবাবগঞ্জের যাত্রী রাত তিনটার পবা হরিপুর বাস চালক নিয়ন্ত্রণ হারালে বাস ধাক্কা খায় বাস সুপারভাইজার যাত্রী শওকত আলী

ফেনীতে কাভার্ড ভ্যান ধাক্কা মোটরসাইকেল নারী নিহত স্বামী মেয়েসহ চারজন আহত ১০টার ফেনী বিসিক শিল্পনগরীর চাড়িপুর ঢাকা চট্টগ্রাম মহাসড়কে দুর্ঘটনা নারী রোকসানা পুলিশ লোকজন কুমিল্লার চৌদ্দগ্রাম আলকরা শিলুরী মোটরসাইকেল চড়ে স্ত্রী সন্তান ফেনী বাসা ফিরছিলেন মাদ্রাসাশিক্ষক আবুল হাসেম বিসিক শিল্পনগরীর চাড়িপুর মহাসড়কে বিপরীত দিক দ্রুতগামী কাভার্ড ভ্যান মোটরসাইকেলের ধাক্কা লাগে মোটরসাইকেলসহ আবুল হাসেম ৪০ স্ত্রী রোকসানা ২৮ সন্তান মাহফুজা ১০ মাহমুদা ৮ মাহবুবা ৪ আহত লোকজন গুরুতর আহত চিকিৎসক রোকসানা আক্তারকে মৃত ঘোষণা আহত ফেনীর মহিপাল কর্মকর্তা আউয়াল ঘটনা পুলিশ

Figure 22: Positive class output of SVM

3.8.1 The Accuracy of SVM

We read 500 news manually to measure the accuracy. We listed the outcomes of classes as follows:

- True Positive (TP): 228
- True Negative (TN): 243
- False positive (FP): 9
- False Negative (FN): 20

Here,

TP: News of road accident classified as road accident

TN: News of other accident classified as other accident

FP: News of other accident classified as road accident

FN: News of road accident classified as other accident

Using the equation for measuring accuracy we found the accuracy of SVM is 94.2%

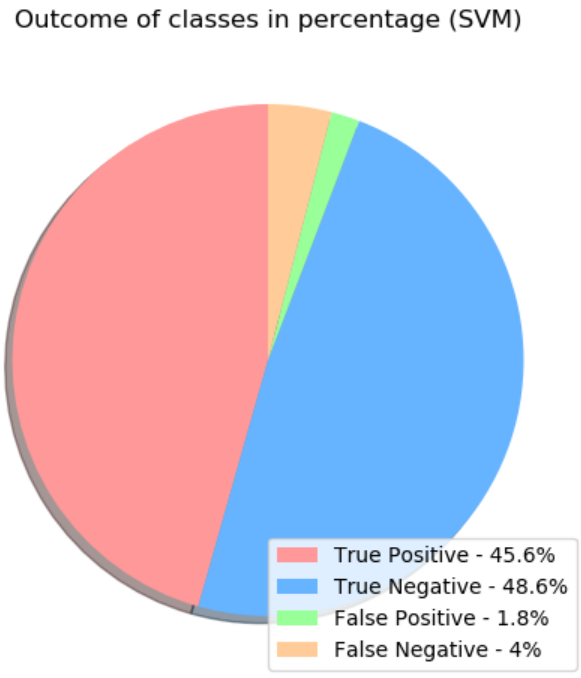


Figure 23: Pie-chart of SVM's outcomes (Classes)

3.9 Extracting Locations

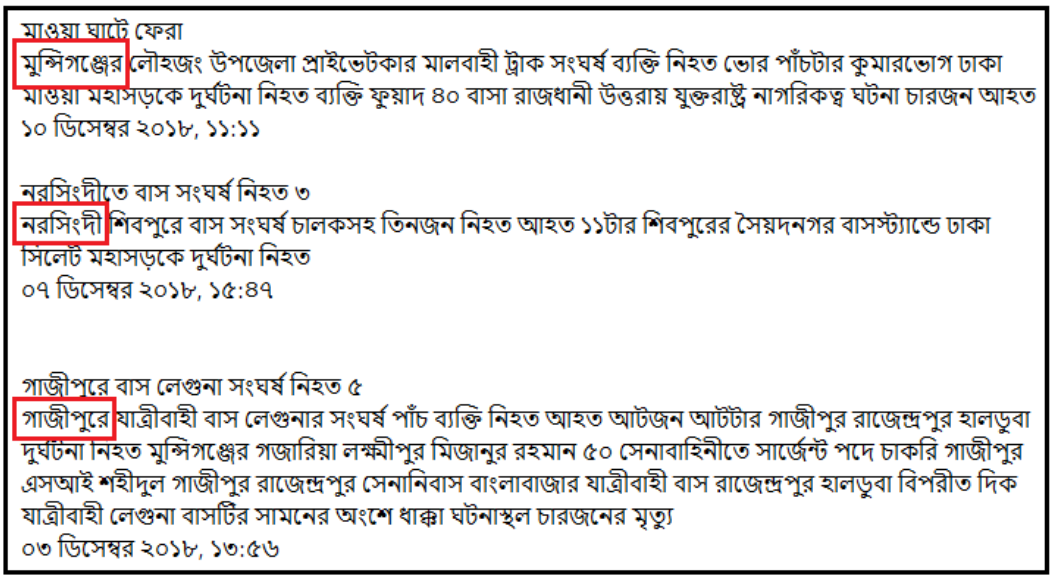
We used F-Beta measure to find which system is more suitable.

We calculated the precision and the recall of the algorithms and selected the value of Beta= 0.5 to give equal importance in precision and recall.

Table 2: Precision, Recall and F-Beta of algorithms

Algorithms	Precision	Recall	F-Beta measure
Naïve Bayes	0.95	**1**	**0.974**
Cosine Similarity	0.89	0.89	0.89
SVM	**0.96**	0.92	0.94

As we can say now Naive Bayes is more suitable so we used the output of Naive Bayes where we kept only accident news with date and headline to extract the place of the accident occurred. After analyzing the news of Prothom-Alo we found more than 95% news start with the name of the location (district) after eliminating stop words (figure-24).

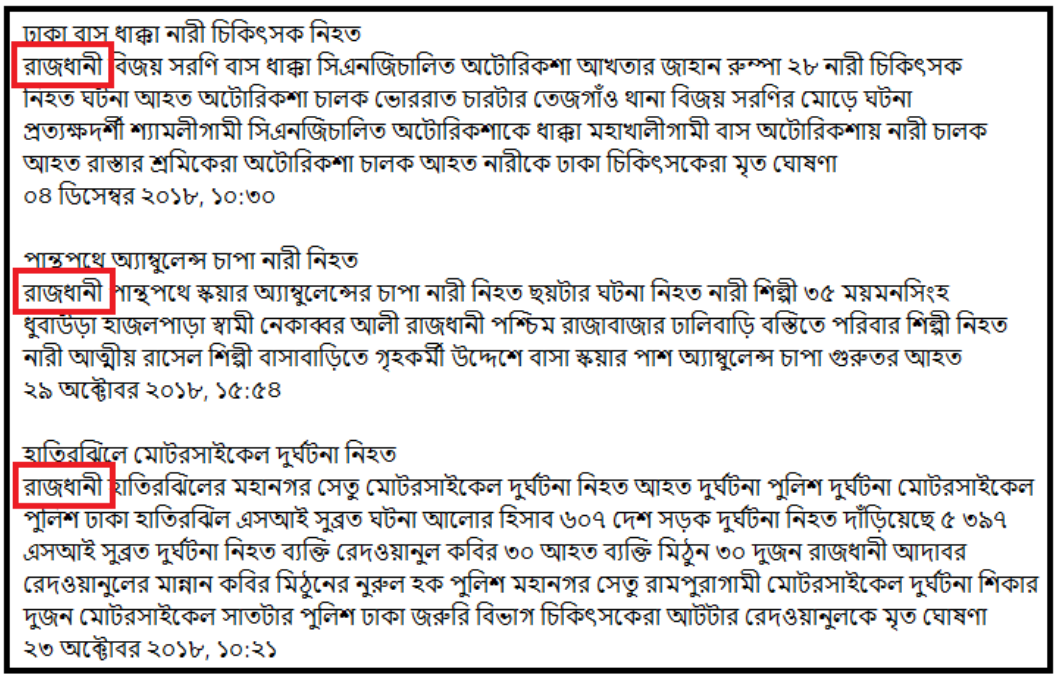

মাওয়া ঘাটে ফেরা
মুন্সিগঞ্জের লৌহজং উপজেলা প্রাইভেটকার মালবাহী ট্রাক সংঘর্ষ ব্যক্তি নিহত ভোর পাঁচটার কুমারভোগ ঢাকা মাওয়া মহাসড়কে দুর্ঘটনা নিহত ব্যক্তি ফুয়াদ ৪০ বাসা রাজধানী উত্তরায় যুক্তরাষ্ট্র নাগরিকত্ব ঘটনা চারজন আহত
১০ ডিসেম্বর ২০১৮, ১১:১১

নরসিংদীতে বাস সংঘর্ষ নিহত ৩
নরসিংদী শিবপুরে বাস সংঘর্ষ চালকসহ তিনজন নিহত আহত ১১টার শিবপুরের সৈয়দনগর বাসস্ট্যান্ডে ঢাকা সিলেট মহাসড়কে দুর্ঘটনা নিহত
০৭ ডিসেম্বর ২০১৮, ১৫:৪৭

গাজীপুরে বাস লেগুনা সংঘর্ষ নিহত ৫
গাজীপুরে যাত্রীবাহী বাস লেগুনার সংঘর্ষ পাঁচ ব্যক্তি নিহত আহত আটজন আটটার গাজীপুর রাজেন্দ্রপুর হালডুবা দুর্ঘটনা নিহত মুন্সিগঞ্জের গজারিয়া লক্ষ্মীপুর মিজানুর রহমান ৫০ সেনাবাহিনীতে সার্জেন্ট পদে চাকরি গাজীপুর এসআই শহীদুল গাজীপুর রাজেন্দ্রপুর সেনানিবাস বাংলাবাজার যাত্রীবাহী বাস রাজেন্দ্রপুর হালডুবা বিপরীত দিক যাত্রীবাহী লেগুনা বাসটির সামনের অংশে ধাক্কা ঘটনাস্থল চারজনের মৃত্যু
০৩ ডিসেম্বর ২০১৮, ১৩:৫৬

Figure 24: Some news from our Naive Bayes Output

But there was a problem, sometimes ঢাকা (Dhaka) is referred to রাজধানী which means capital (fig-25).


ঢাকা বাস ধাক্কা নারী চিকিৎসক নিহত
রাজধানী বিজয় সরণি বাস ধাক্কা সিএনজিচালিত অটোরিকশা আখতার জাহান রুম্পা ২৮ নারী চিকিৎসক নিহত ঘটনা আহত অটোরিকশা চালক ভোররাত চারটার তেজগাঁও থানা বিজয় সরণির মোড়ে ঘটনা প্রত্যক্ষদর্শী শ্যামলীগামী সিএনজিচালিত অটোরিকশাকে ধাক্কা মহাখালীগামী বাস অটোরিকশায় নারী চালক আহত রাস্তার শ্রমিকেরা অটোরিকশা চালক আহত নারীকে ঢাকা চিকিৎসকেরা মৃত ঘোষণা
০৪ ডিসেম্বর ২০১৮, ১০:৩০

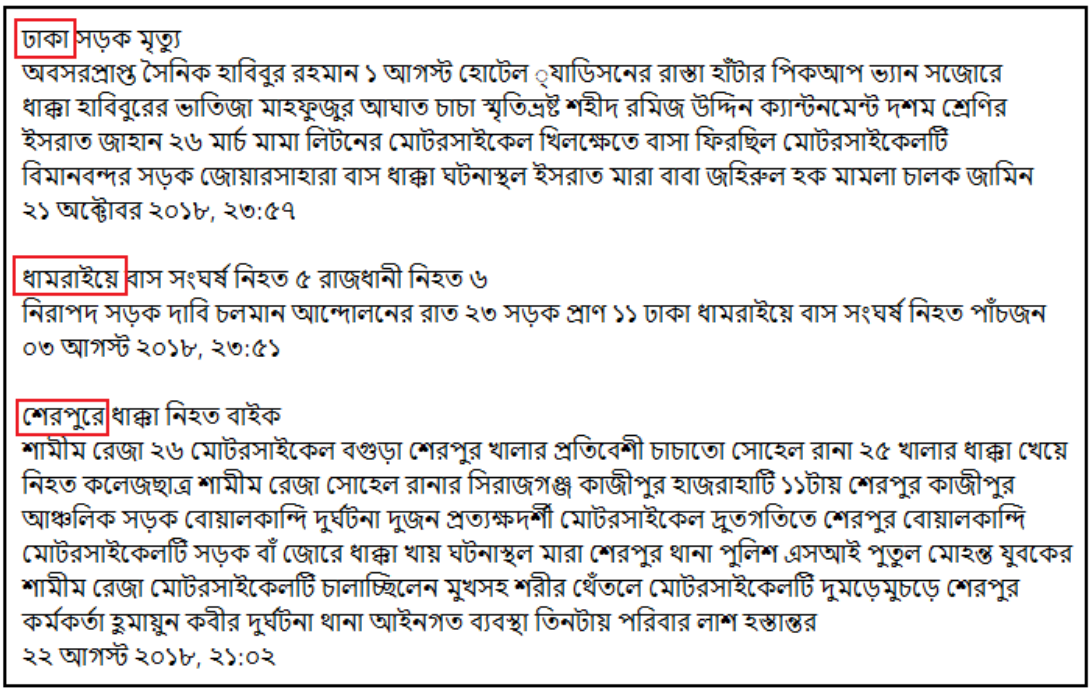
পান্থপথে অ্যাম্বুলেন্স চাপা নারী নিহত
রাজধানী পান্থপথে স্কয়ার অ্যাম্বুলেন্সের চাপা নারী নিহত ছয়টার ঘটনা নিহত নারী শিল্পী ৩৫ ময়মনসিংহ ধুবাউড়া হাজলপাড়া স্বামী নেকাব্বর আলী রাজধানী পশ্চিম রাজাবাজার ঢালিবাড়ি বস্তিতে পরিবার শিল্পী নিহত নারী আত্মীয় রাসেল শিল্পী বাসাবাড়িতে গৃহকর্মী উদ্দেশে বাসা স্কয়ার পাশ অ্যাম্বুলেন্স চাপা গুরুতর আহত
২৯ অক্টোবর ২০১৮, ১৫:৫৪

হাতিরঝিলে মোটরসাইকেল দুর্ঘটনা নিহত
রাজধানী হাতিরঝিলের মহানগর সেতু মোটরসাইকেল দুর্ঘটনা নিহত আহত দুর্ঘটনা পুলিশ দুর্ঘটনা মোটরসাইকেল পুলিশ ঢাকা হাতিরঝিল এসআই সুব্রত ঘটনা আলোর হিসাব ৬০৭ দেশ সড়ক দুর্ঘটনা নিহত দাঁড়িয়েছে ৫ ৩৯৭ এসআই সুব্রত দুর্ঘটনা নিহত ব্যক্তি রেদওয়ানুল কবির ৩০ আহত ব্যক্তি মিঠুন ৩০ দুজন রাজধানী আদাবর রেদওয়ানুলের মান্নান কবির মিঠুনের নুরুল হক পুলিশ মহানগর সেতু রামপুরাগামী মোটরসাইকেল দুর্ঘটনা শিকার দুজন মোটরসাইকেল সাতটার পুলিশ ঢাকা জরুরি বিভাগ চিকিৎসকেরা আটটার রেদওয়ানুলকে মৃত ঘোষণা
২৩ অক্টোবর ২০১৮, ১০:২১

Figure 25: Some news that use Dhaka as Rajdhani

So whenever we found রাজধানী we consider it as ঢাকা.

After using the first pattern, we got some news which doesn't have district name in the first word of the news body. But some of the news carries the district name on the first word of the heading. We matched the pattern also and found a lot of the news matched to it.



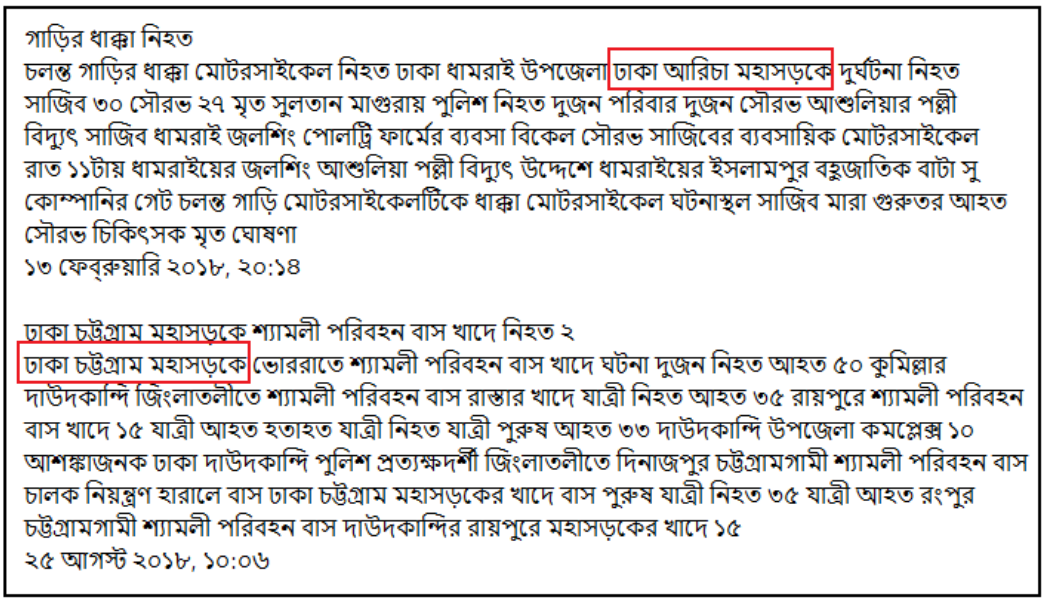
ঢাকা সড়ক মৃত্যু
অবসরপ্রাপ্ত সৈনিক হাবিবুর রহমান ১ আগস্ট হোটেল ্যাডিসনের রাস্তা হাঁটার পিকআপ ভ্যান সজোরে ধাক্কা হাবিবুরের ভাতিজা মাহফুজুর আঘাত চাচা স্মৃতিভ্রষ্ট শহীদ রমিজ উদ্দিন ক্যান্টনমেন্ট দশম শ্রেণির ইসরাত জাহান ২৬ মার্চ মামা লিটনের মোটরসাইকেল খিলক্ষেতে বাসা ফিরছিল মোটরসাইকেলটি বিমানবন্দর সড়ক জোয়ারসাহারা বাস ধাক্কা ঘটনাস্থল ইসরাত মারা বাবা জহিরুল হক মামলা চালক জামিন
২১ অক্টোবর ২০১৮, ২৩:৫৭

ধামরাইয়ে বাস সংঘর্ষ নিহত ৫ রাজধানী নিহত ৬
নিরাপদ সড়ক দাবি চলমান আন্দোলনের রাত ২৩ সড়ক প্রাণ ১১ ঢাকা ধামরাইয়ে বাস সংঘর্ষ নিহত পাঁচজন
০৩ আগস্ট ২০১৮, ২৩:৫১

শেরপুরে ধাক্কা নিহত বাইক
শামীম রেজা ২৬ মোটরসাইকেল বগুড়া শেরপুর খালার প্রতিবেশী চাচাতো সোহেল রানা ২৫ খালার ধাক্কা খেয়ে নিহত কলেজছাত্র শামীম রেজা সোহেল রানার সিরাজগঞ্জ কাজীপুর হাজরাহাটি ১১টায় শেরপুর কাজীপুর আঞ্চলিক সড়ক বোয়ালকান্দি দুর্ঘটনা দুজন প্রত্যক্ষদর্শী মোটরসাইকেল দ্রুতগতিতে শেরপুর বোয়ালকান্দি মোটরসাইকেলটি সড়ক বাঁ জোরে ধাক্কা খায় ঘটনাস্থল মারা শেরপুর থানা পুলিশ এসআই পুতুল মোহন্ত যুবকের শামীম রেজা মোটরসাইকেলটি চালাচ্ছিলেন মুখসহ শরীর থেঁতলে মোটরসাইকেলটি দুমড়েমুচড়ে শেরপুর কর্মকর্তা হুমায়ুন কবীর দুর্ঘটনা থানা আইনগত ব্যবস্থা তিনটায় পরিবার লাশ হস্তান্তর
২২ আগস্ট ২০১৮, ২১:০২

Figure 26: Some news where accident location is the first word of the heading.

After using the second one, we still had some news which doesn't have district name at the first word of the news body and also doesn't have district name at the first word of the news heading. We noticed that some accidents occur on the highway. In these news "Prothom-Alo" maintains a pattern like– "ঢাকা আরিচা মহাসড়কে". In that type of news, we considered the first occurring place name as the accident location. So here "ঢাকা" is considered as the accident place.



গাড়ির ধাক্কা নিহত
চলন্ত গাড়ির ধাক্কা মোটরসাইকেল নিহত ঢাকা ধামরাই উপজেলা ঢাকা আরিচা মহাসড়কে দুর্ঘটনা নিহত সাজিব ৩০ সৌরভ ২৭ মৃত সুলতান মাগুরায় পুলিশ নিহত দুজন পরিবার দুজন সৌরভ আশুলিয়ার পল্লী বিদ্যুৎ সাজিব ধামরাই জলশিং পোলট্রি ফার্মের ব্যবসা বিকেল সৌরভ সাজিবের ব্যবসায়িক মোটরসাইকেল রাত ১১টায় ধামরাইয়ের জলশিং আশুলিয়া পল্লী বিদ্যুৎ উদ্দেশে ধামরাইয়ের ইসলামপুর বহুজাতিক বাটা সু কোম্পানির গেট চলন্ত গাড়ি মোটরসাইকেলটিকে ধাক্কা মোটরসাইকেল ঘটনাস্থল সাজিব মারা গুরুতর আহত সৌরভ চিকিৎসক মৃত ঘোষণা
১৩ ফেব্রুয়ারি ২০১৮, ২০:১৪

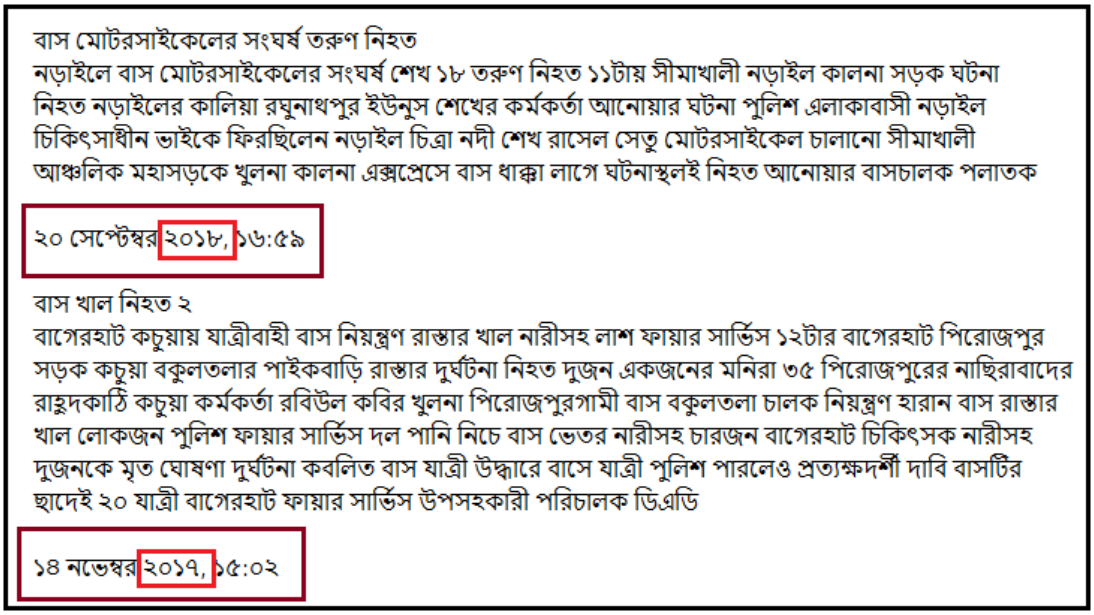
ঢাকা চট্টগ্রাম মহাসড়কে শ্যামলী পরিবহন বাস খাদে নিহত ২
ঢাকা চট্টগ্রাম মহাসড়কে ভোররাতে শ্যামলী পরিবহন বাস খাদে ঘটনা দুজন নিহত আহত ৫০ কুমিল্লার দাউদকান্দি জিংলাতলীতে শ্যামলী পরিবহন বাস রাস্তার খাদে যাত্রী নিহত আহত ৩৫ রায়পুরে শ্যামলী পরিবহন বাস খাদে ১৫ যাত্রী আহত হতাহত যাত্রী নিহত যাত্রী পুরুষ আহত ৩৩ দাউদকান্দি উপজেলা কমপ্লেক্স ১০ আশঙ্কাজনক ঢাকা দাউদকান্দি পুলিশ প্রত্যক্ষদর্শী জিংলাতলীতে দিনাজপুর চট্টগ্রামগামী শ্যামলী পরিবহন বাস চালক নিয়ন্ত্রণ হারালে বাস ঢাকা চট্টগ্রাম মহাসড়কের খাদে বাস পুরুষ যাত্রী নিহত ৩৫ যাত্রী আহত রংপুর চট্টগ্রামগামী শ্যামলী পরিবহন বাস দাউদকান্দির রায়পুরে মহাসড়কের খাদে ১৫
২৫ আগস্ট ২০১৮, ১০:০৬

Figure 27: Some news from our final data

3.10 Extracting Year

After the news crawling we have a date for each of the news and all the date are showing the same pattern, like- "২০ অক্টোবর ২০১৮, ১৮: ৪১"

We figured out that, the accident year is the 3rd word of the date line. So, we split it and took the 3rd word as our accident occurring date.



বাস মোটরসাইকেলের সংঘর্ষ তরুণ নিহত
নড়াইলে বাস মোটরসাইকেলের সংঘর্ষ শেখ ১৮ তরুণ নিহত ১১টায় সীমাখালী নড়াইল কালনা সড়ক ঘটনা নিহত নড়াইলের কালিয়া রঘুনাথপুর ইউনুস শেখের কর্মকর্তা আনোয়ার ঘটনা পুলিশ এলাকাবাসী নড়াইল চিকিৎসাধীন ভাইকে ফিরছিলেন নড়াইল চিত্রা নদী শেখ রাসেল সেতু মোটরসাইকেল চালানো সীমাখালী আঞ্চলিক মহাসড়কে খুলনা কালনা এক্সপ্রেসে বাস ধাক্কা লাগে ঘটনাস্থলই নিহত আনোয়ার বাসচালক পলাতক

২০ সেপ্টেম্বর ২০১৮, ১৬:৫৯

বাস খাল নিহত ২
বাগেরহাট কচুয়ায় যাত্রীবাহী বাস নিয়ন্ত্রণ রাস্তার খাল নারীসহ লাশ ফায়ার সার্ভিস ১২টার বাগেরহাট পিরোজপুর সড়ক কচুয়া বকুলতলার পাইকবাড়ি রাস্তার দুর্ঘটনা নিহত দুজন একজনের মনিরা ৩৫ পিরোজপুরের নাছিরাবাদের রাহুদকাঠি কচুয়া কর্মকর্তা রবিউল কবির খুলনা পিরোজপুরগামী বাস বকুলতলা চালক নিয়ন্ত্রণ হারান বাস রাস্তার খাল লোকজন পুলিশ ফায়ার সার্ভিস দল পানি নিচে বাস ভেতর নারীসহ চারজন বাগেরহাট চিকিৎসক নারীসহ দুজনকে মৃত ঘোষণা দুর্ঘটনা কবলিত বাস যাত্রী উদ্ধারে বাসে যাত্রী পুলিশ পারলেও প্রত্যক্ষদর্শী দাবি বাসটির ছাদেই ২০ যাত্রী বাগেরহাট ফায়ার সার্ভিস উপসহকারী পরিচালক ডিএডি

১৪ নভেম্বর ২০১৭, ১৫:০২

Figure 28: Some news from our final data

3.11 Accident Rates in Divisions

Using the data we got after grouping the results into years, we made diagrams where we showed our result, division wise which contains all the districts of the division.

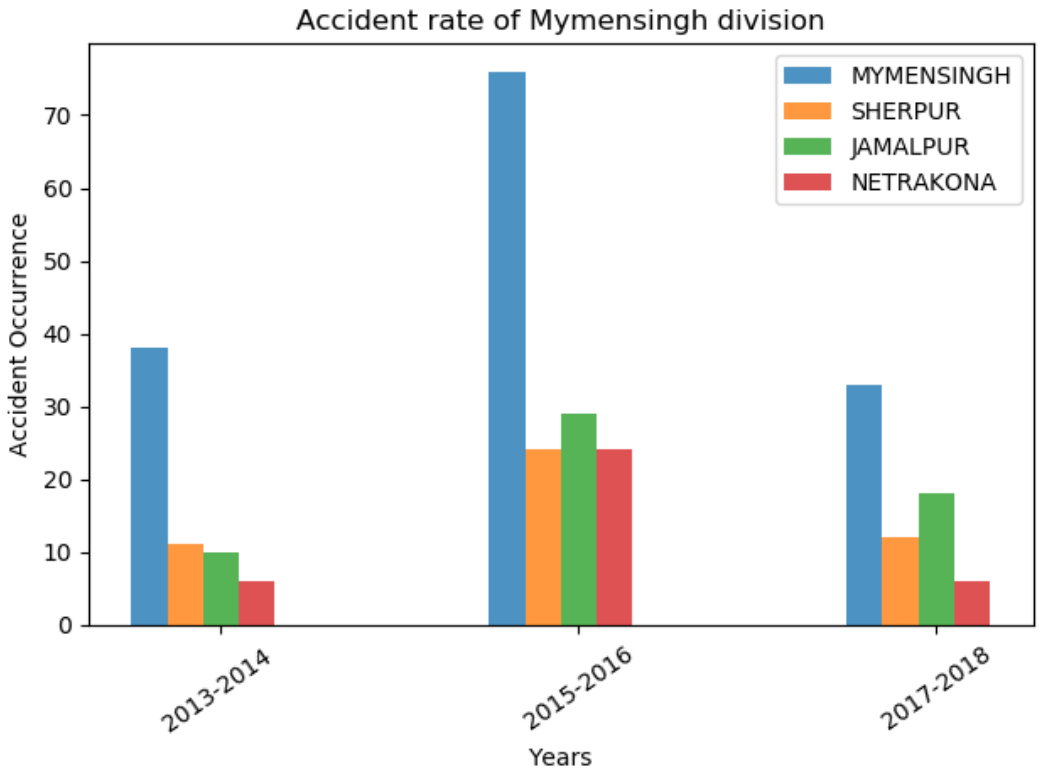


Figure 29: Bar chart of accident rate (Mymensingh division)

Figure-29 illustrates the accident rate of all districts of Mymensingh division from 2013 to 2018. Compare to all of the districts of Mymensingh division, Mymensingh district is the most road accident occurred place and Jamalpur is the second highest accident occurring place in this division.

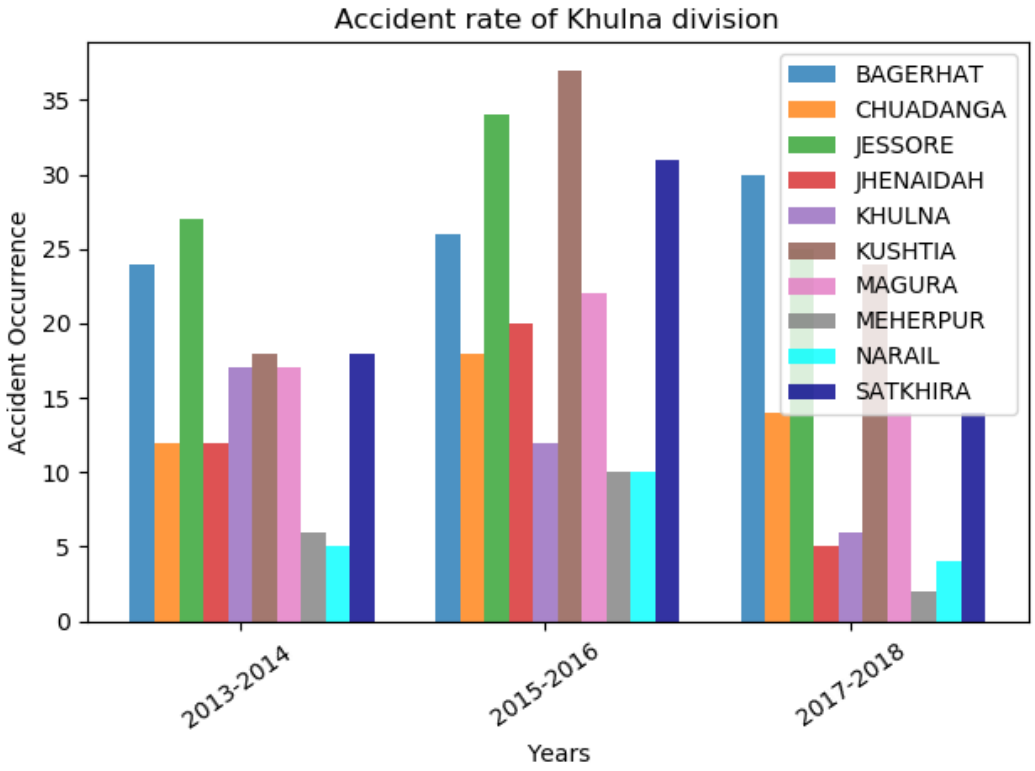


Figure 30: Bar chart of accident rate (Khulna division)

Figure-30 illustrates the accident rate of all districts of Khulna division from 2013 to 2018. Compare to all of the districts of Khulna division, it is evident that Kushtia is the most accident occurring district and Meherpur is the less accident occurring district.

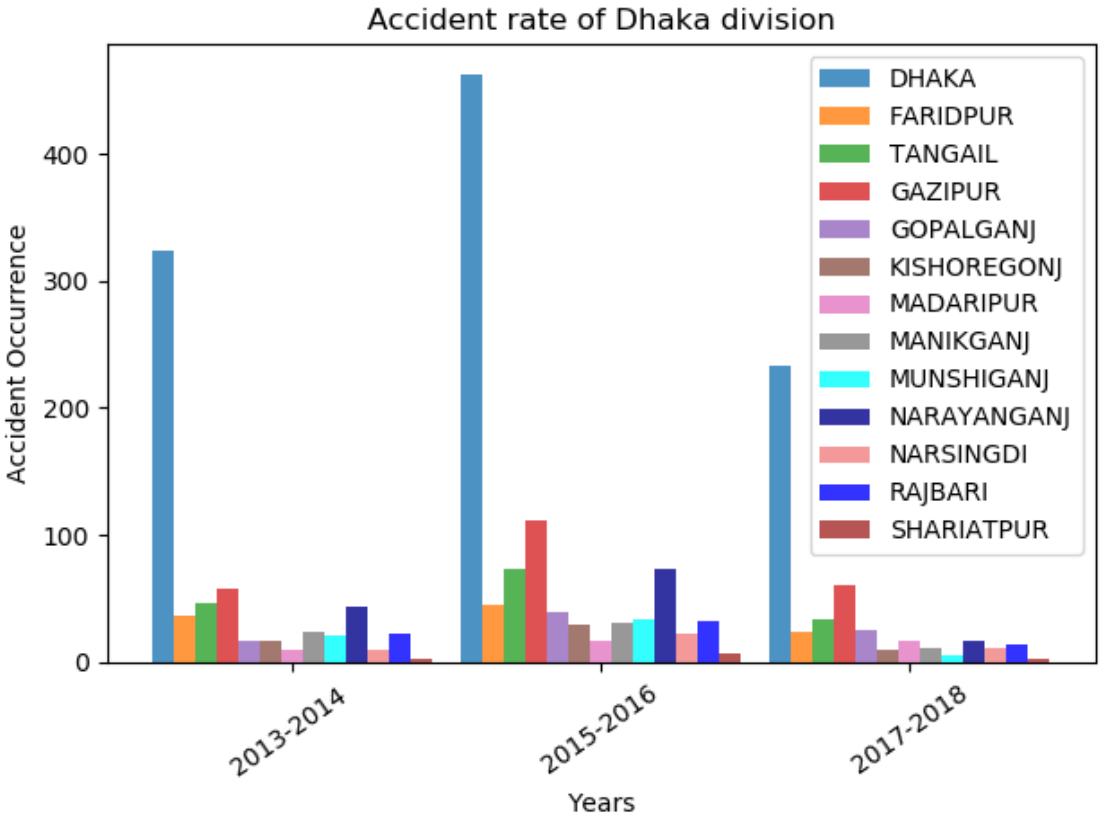


Figure 31: Bar chart of accident rate (Dhaka division)

Figure-31 illustrates the accident rate of all districts of Dhaka division from 2013 to 2018. Compare to all of the districts of Dhaka division, Dhaka district is the most road accident occurred place and Gazipur is the second highest accident occurring place in this division. Accident rate of the year 2015-2016 is higher than 2013-2014 and 2017-2018.

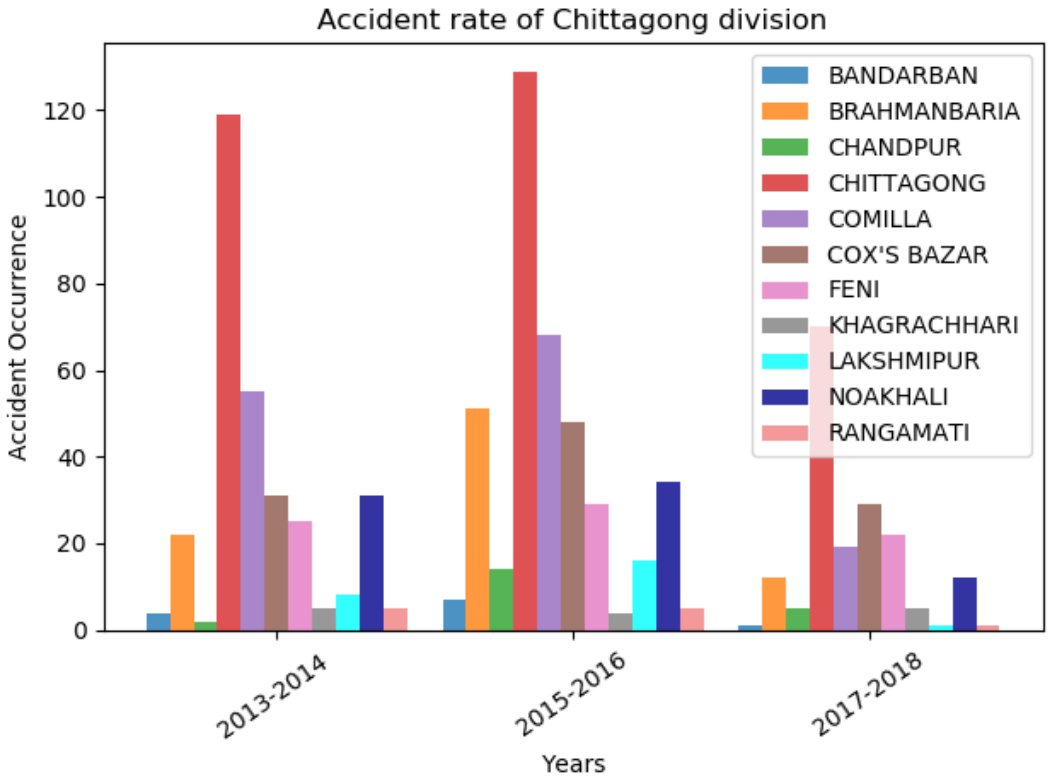


Figure 32: Bar chart of accident rate (Chittagong division)

Figure-32 illustrates the accident rate of all districts of Chittagong division from 2013 to 2018. Compare to all of the districts of Chittagong division Chittagong district is the most road accident occurring place. Comilla is the second highest road accident occurring place in this division at 2013-2014 and 2015-2016. But Cox's Bazar was at the second position in 2017-2018. And all of the accident rates of the year 2015-2016 are higher than 2013-2014 and 2017-2018 in this Division.

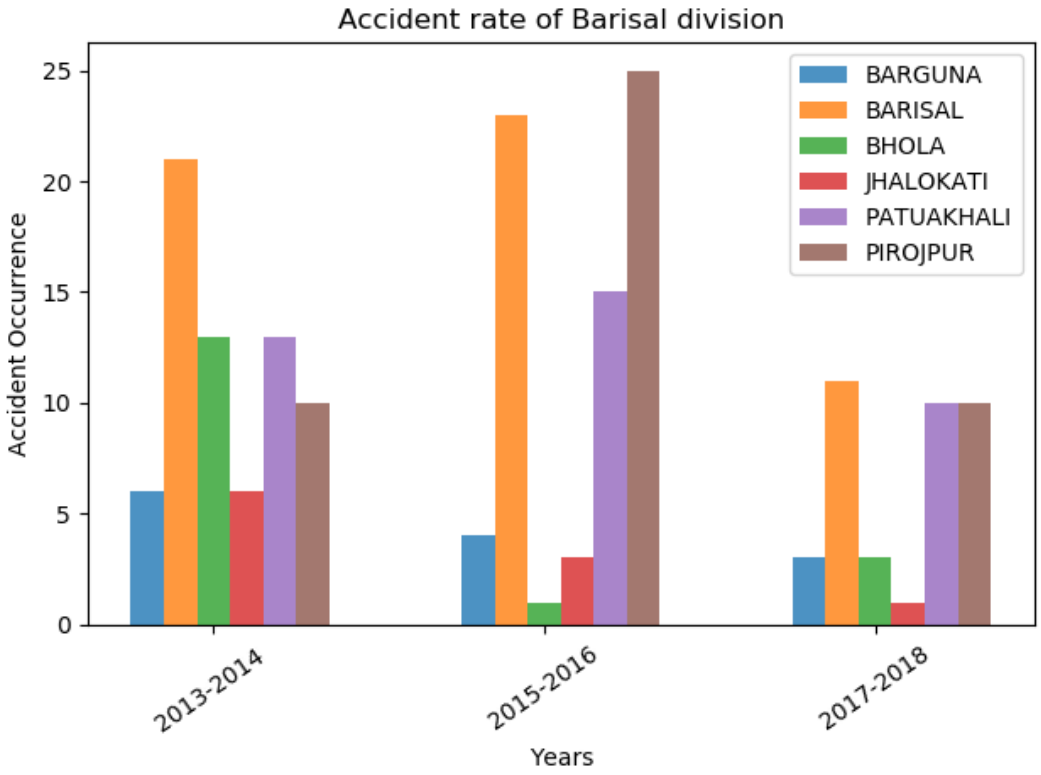


Figure 33: Bar chart of accident rate (Barisal division)

Figure-33 illustrates the accident rate of all districts of Barisal division from 2013 to 2018. Compare to all of the districts of Barisal division, Barisal district is the most road accident occurring place in between 2013-2014 and 2017-2018. But in between 2015-2016, Pirojpur becomes the most road accident occurring district of Barisal division.

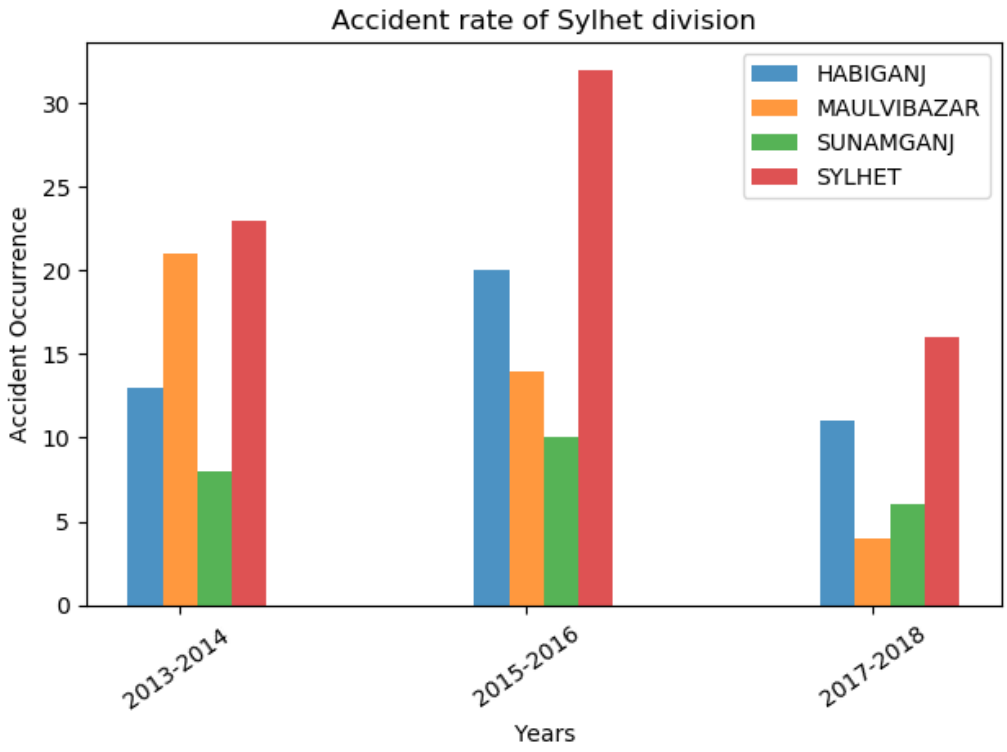


Figure 34: Bar chart of accident rate (Sylhet division)

Figure-34 shows the accident rate of the Sylhet division. Here, The Sylhet district is the most accident occurring district of Sylhet division. We see the Sunamgonj district is the less accident occurring district of Sylhet division.

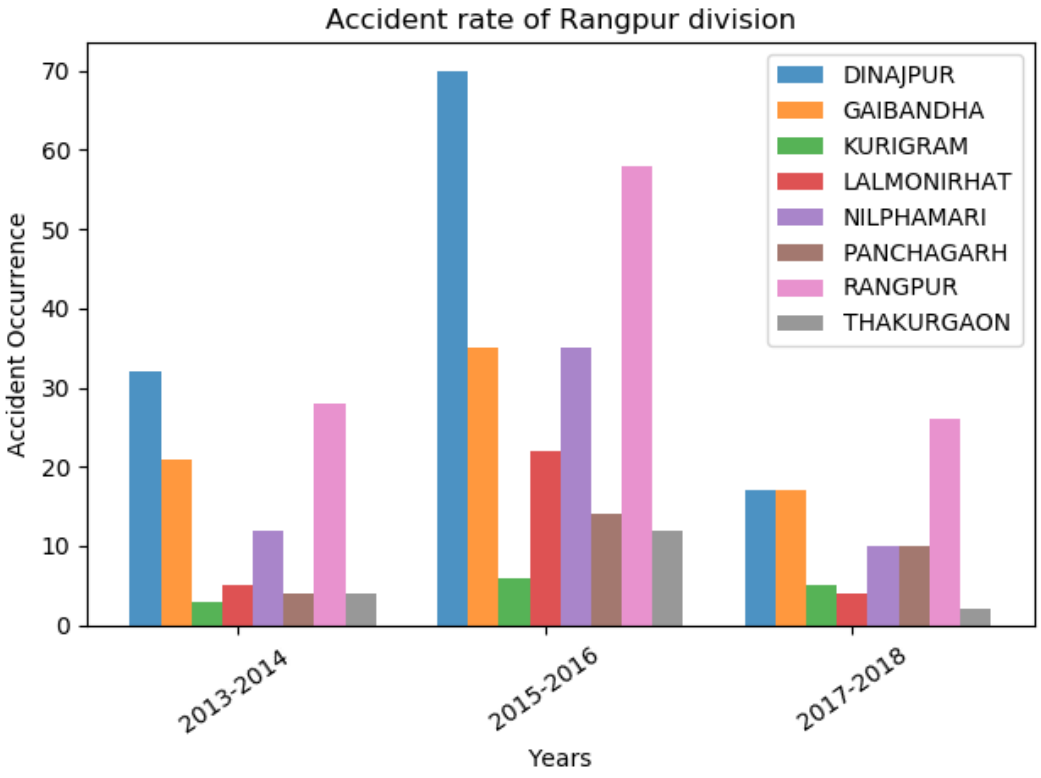


Figure 35: Bar chart of accident rate (Rangpur division)

Figure-35 illustrates the accident rate of all districts of Rangpur division from 2013 to 2018. Compare to all of the districts of Rangpur division, Dinajpur district is the most road accident occurring place in between 2013-2014 and 2015-2016. But in between 2017-2018, Rangpur becomes the most road accident occurring district of Rangpur division.

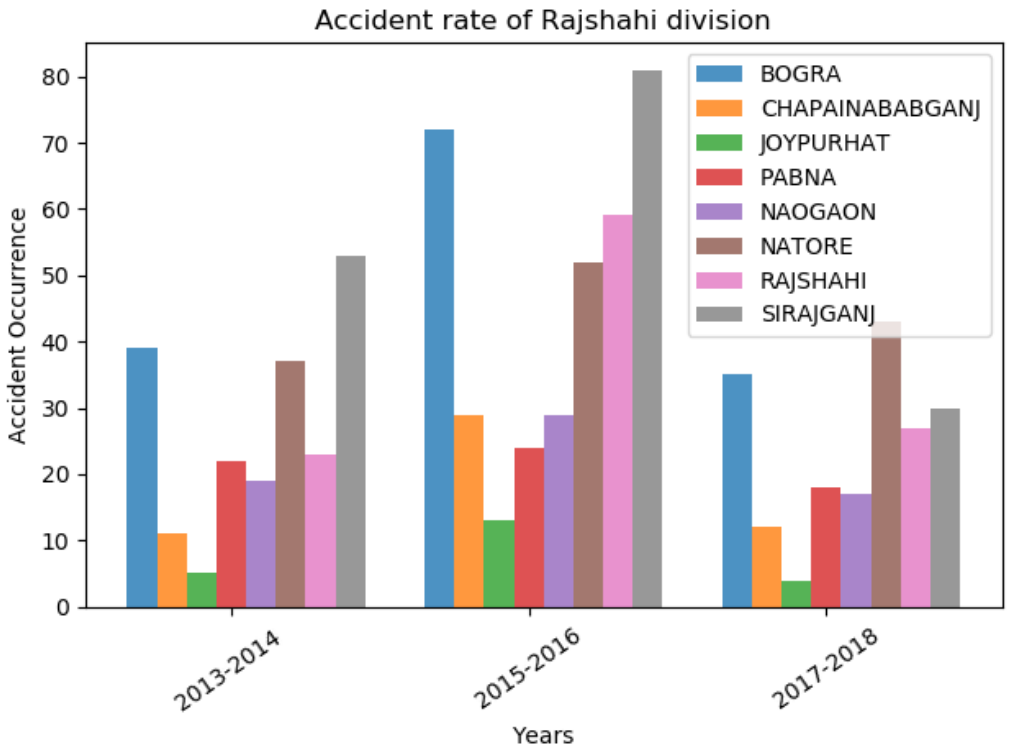


Figure 36: Bar chart of accident rate (Rajshahi division)

Figure-36 illustrates the accident rate of all districts of Rajshahi division from 2013 to 2018. Compare to all of the districts of Rajshahi division, Sirajganj district is the most road accident occurring place in between 2013-2014 and 2015-2016. But in between 2017-2018, Natore becomes the most road accident occurring district of Rajshahi division. We also see that Joypurhat district is the less accident occurring district of Rajshahi division.

3.12 Heatmap

We collected all the districts name and the number of road accidents occurred in each of the districts from 2013 to 2018. Then to view it on a map, we used Google Map-API. We specified the latitude and longitude of a particular district as location and the number of road accident occurred there as weight. Then Google map-API plotted a circle at the location with respect to the weight.

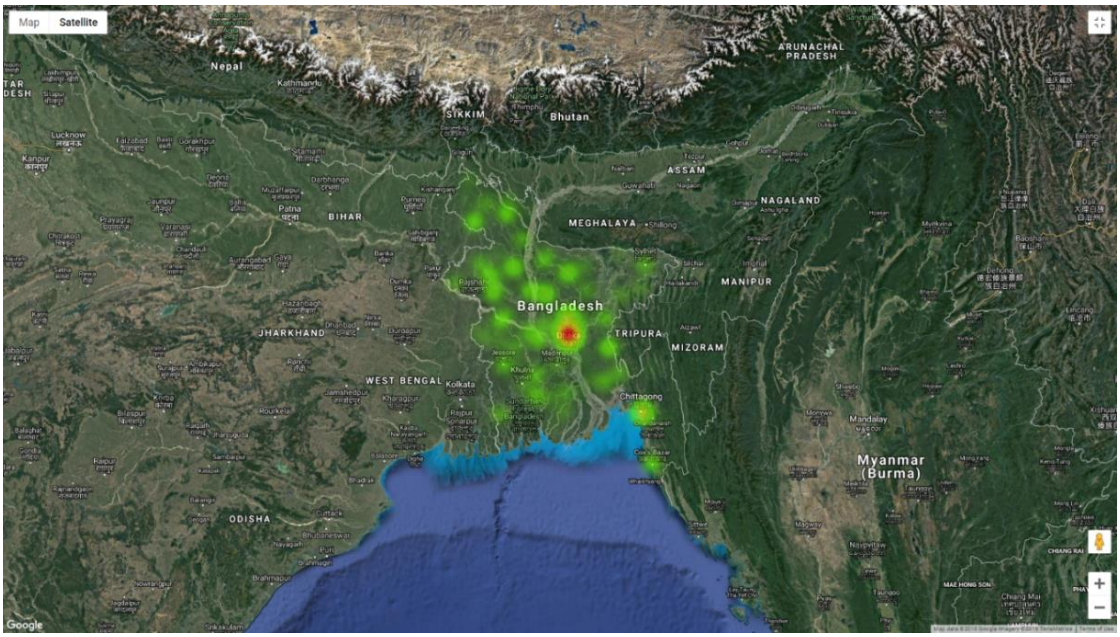


Figure 37: Satellite of view of Bangladesh

For a better view and to identify the districts name, we zoomed in the map:

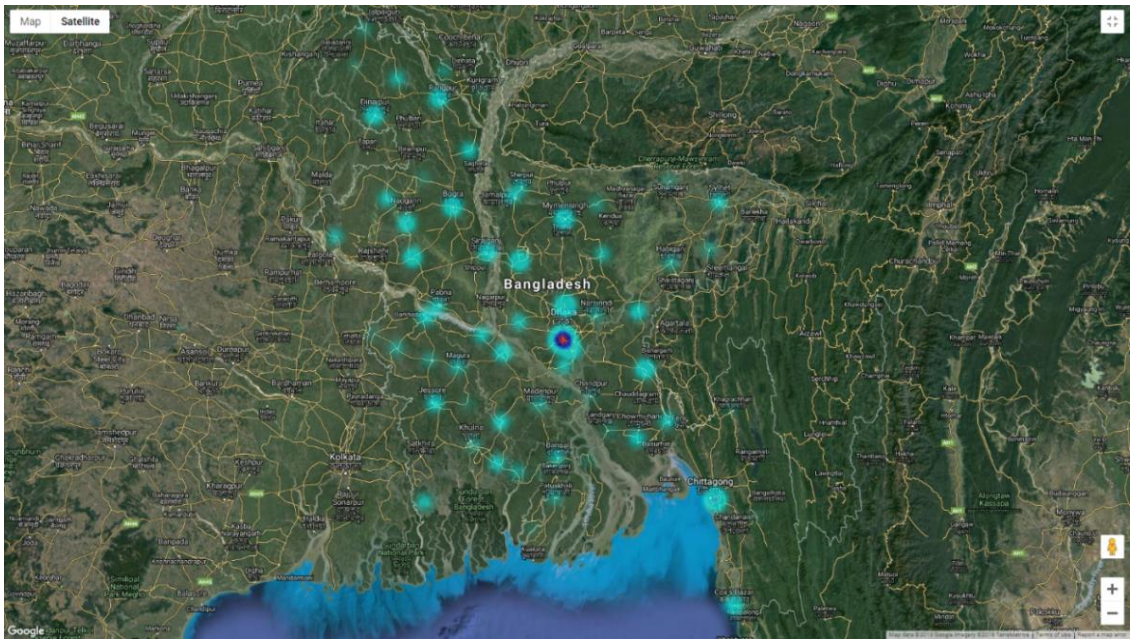


Figure 38: Satellite of view (zoom) of Bangladesh

CHAPTER 4

Result and Discussion

4.1 Vehicle Observation

We finally tried to detect the vehicle involved with accidents. To do this, we made a list of possible vehicles by reading almost 1500 news.

We wrote code to find the vehicle name from accidents news. Sometimes, there was only one vehicle name in the news, and sometimes there was multiple vehicles name.

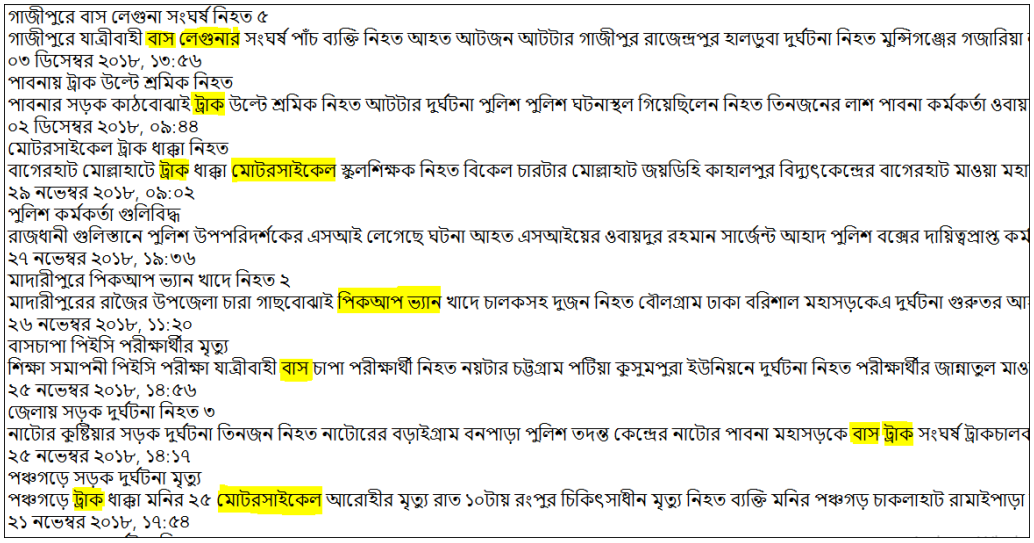

গাজীপুরে বাস লেগুনা সংঘর্ষ নিহত ৫
গাজীপুরে যাত্রীবাহী বাস লেগুনার সংঘর্ষ পাঁচ ব্যক্তি নিহত আহত আটজন আটটার গাজীপুর রাজেন্দ্রপুর হালডুবা দুর্ঘটনা নিহত মুন্সিগঞ্জের গজারিয়া
০৩ ডিসেম্বর ২০১৮, ১৩:৫৬
পাবনায় ট্রাক উল্টে শ্রমিক নিহত
পাবনার সড়ক কাঠবোঝাই ট্রাক উল্টে শ্রমিক নিহত আটটার দুর্ঘটনা পুলিশ পুলিশ ঘটনাস্থল গিয়েছিলেন নিহত তিনজনের লাশ পাবনা কর্মকর্তা ওবায়
০২ ডিসেম্বর ২০১৮, ০৯:৪৪
মোটরসাইকেল ট্রাক ধাক্কা নিহত
বাগেরহাট মোল্লাহাটে ট্রাক ধাক্কা মোটরসাইকেল স্কুলশিক্ষক নিহত বিকেল চারটার মোল্লাহাট জয়ডিহি কাহালপুর বিদ্যুৎকেন্দ্রের বাগেরহাট মাওয়া মহা
২৯ নভেম্বর ২০১৮, ০৯:০২
পুলিশ কর্মকর্তা গুলিবিদ্ধ
রাজধানী গুলিস্তানে পুলিশ উপপরিদর্শকের এসআই লেগেছে ঘটনা আহত এসআইয়ের ওবায়দুর রহমান সার্জেন্ট আহাদ পুলিশ বক্সের দায়িত্বপ্রাপ্ত কম
২৭ নভেম্বর ২০১৮, ১৯:৩৬
মাদারীপুরে পিকআপ ভ্যান খাদে নিহত ২
মাদারীপুরের রাজৈর উপজেলা চারা গাছবোঝাই পিকআপ ভ্যান খাদে চালকসহ দুজন নিহত বৌলগ্রাম ঢাকা বরিশাল মহাসড়কেএ দুর্ঘটনা গুরুতর আ
২৬ নভেম্বর ২০১৮, ১১:২০
বাসচাপা পিইসি পরীক্ষার্থীর মৃত্যু
শিক্ষা সমাপনী পিইসি পরীক্ষা যাত্রীবাহী বাস চাপা পরীক্ষার্থী নিহত নয়টার চট্টগ্রাম পটিয়া কুসুমপুরা ইউনিয়নে দুর্ঘটনা নিহত পরীক্ষার্থীর জান্নাতুল মাও
২৫ নভেম্বর ২০১৮, ১৪:৫৬
জেলায় সড়ক দুর্ঘটনা নিহত ৩
নাটোর কুষ্টিয়ার সড়ক দুর্ঘটনা তিনজন নিহত নাটোরের বড়াইগ্রাম বনপাড়া পুলিশ তদন্ত কেন্দ্রের নাটোর পাবনা মহাসড়কে বাস ট্রাক সংঘর্ষ ট্রাকচালব
২৫ নভেম্বর ২০১৮, ১৪:১৭
পঞ্চগড়ে সড়ক দুর্ঘটনা মৃত্যু
পঞ্চগড়ে ট্রাক ধাক্কা মনির ২৫ মোটরসাইকেল আরোহীর মৃত্যু রাত ১০টায় রংপুর চিকিৎসাধীন মৃত্যু নিহত ব্যক্তি মনির পঞ্চগড় চাকলাহাট রামাইপাড়া
২১ নভেম্বর ২০১৮, ১৭:৫৪

Figure 39: Single and multiple vehicles in an accident

For a single vehicle involved in a road accident news, we identified that vehicle as the accident vehicle. And for multiple vehicles, we identified each of these as the accident vehicle. From 5939 road accident news, we found the following list:

Table 3: Accident rate of different vehicles

Vehicles Names	Number of accidents	Accident rate
BUS	1969	36.46%
TRUCK	1489	27.57%
MICRO-BUS	237	4.38%
MOTOR CYCLE	675	12.5%
CNG	28	0.5%
RICKSHAW	368	7.9%
LAGUNA	26	0.5%
PICKUP-VAN	277	5.1%
CAVARD VAN	201	3.7%
CHANDER GARI	11	0.2%
VOTVOTI	64	1.2%

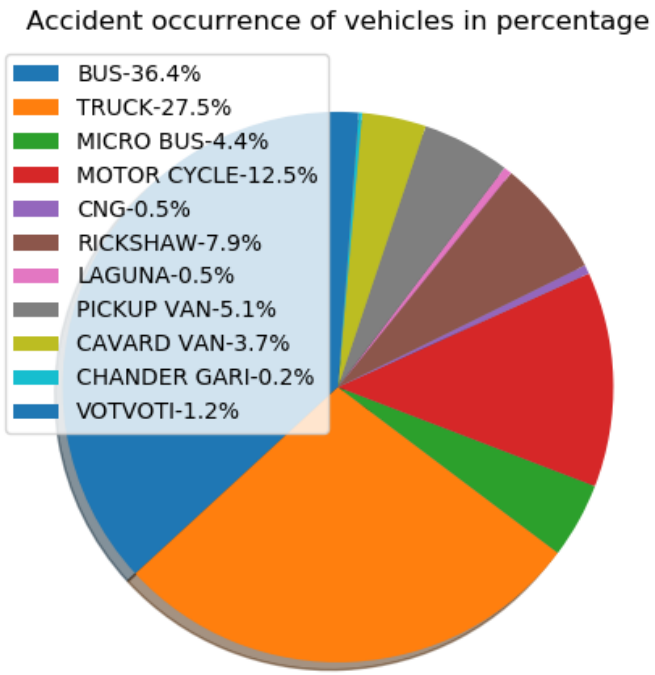


Figure 40: A pie-chart of the accident occurring based on vehicle

Figure-40 illustrates that the Buses and the Trucks are mostly involved with road accidents in Bangladesh. In total 63.9% of the total road accidents are caused by Buses and Trucks (fig-40).

4.2 Performance measuring

We created a graph where we showed comparisons between Naive Bayes, Cosine Similarity, Support Vector Machine.

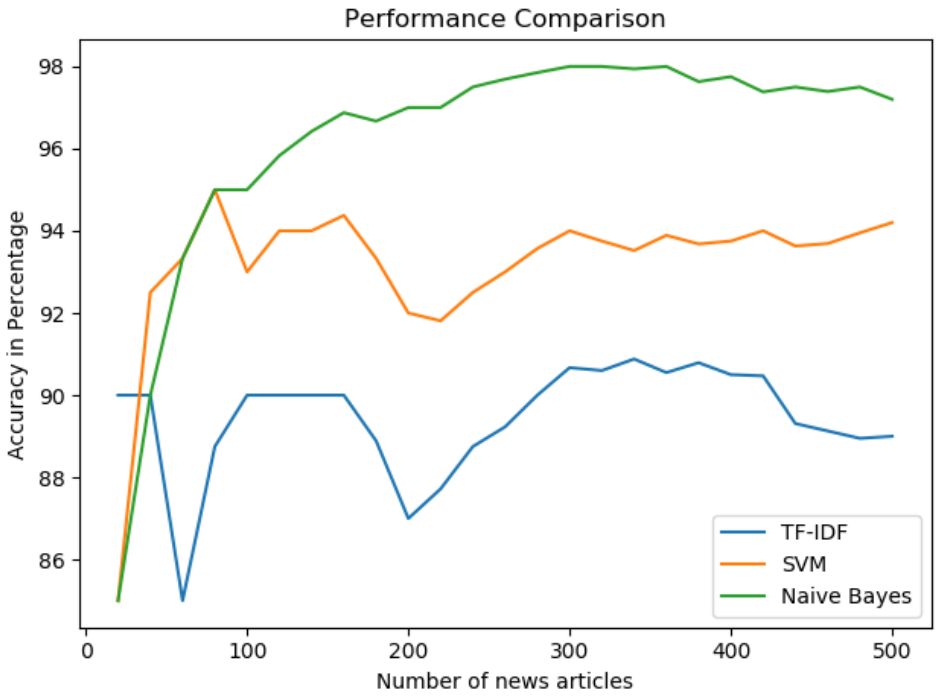


Figure 41: Performance comparison between Naive Bayes, Cosine similarity and SVM

From Figure-41, we can clearly see that the performance of Naive Bayes increased with the data increase and its performance is higher than the rest of two. The performance of SVM is in between Naive Bayes and Cosine Similarity, higher than cosine similarity but less than Naïve Bayes. And the Cosine Similarity has the lowest performance among these three.

CHAPTER 5

Conclusion

To reduce the number of Road Accidents and ensuring a safe journey system, it's required to know where road accidents occur often and which is the most accident occurring vehicle. Unfortunately, there haven't much previous work been done to know the often road accident location retrieving news from different Bengali online newspapers. Though our developed approach is not sufficient but we expect that will surely be helpful for the government, general people, drivers, police or tourists to be more careful about the most occurring accident places and vehicles and taking necessary steps to make the daily journey safer.

5.1 Limitations

- We have crawled news from only one newspaper. Which name was "Prothom-Alo". Which may not have news of all regions. If we could have used regional newspapers as a resource we might have got better accuracy.
- Our work is only districts and division based.
- We used five patterns to extract the accident location which almost covered 90% of news of "Prothom-Alo". But for other newspapers, we may need to find some other patterns.

5.2 Future Work

- The whole process can be done dynamically. In one go, the whole process of the work (Crawling, Parsing, Converting, Clustering, Extracting Locations, Showing statistics, plotting at Heat-map etc.) can be done. In other words- Automating the process.
- Finding the number of casualties by particular accidents.
- Using more newspapers as a resource.
- We will work with Sub-districts along with districts and divisions.

References

1. Monica Peshave.How search engines work and a web crawler application.
2. Heydon, A. & Najork, M. World Wide Web (1999) 2: 219. https://doi.org/10.1023/A:1019213109274
3. https://www.researchgate.net/figure/Typical-high-level-architecture-of-a-Web-crawler-involving-a-scheduler-and-a_fig11_220466525
4. Rajib Chandra Das, Md. Ruhul Amin and Md. Jumman Hossain. Predicting crime occurrence retrieving news from different online newspapers.
5. Zhang K., Xu H., Tang J., Li J. (2006) Keyword Extraction Using Support Vector Machine. In: Yu J.X., Kitsuregawa M., Leong H.V. (eds) Advances in Web-Age Information Management. WAIM 2006. Lecture Notes in Computer Science, vol 4016. Springer, Berlin, Heidelberg
6. Introduction to Machine Learning, Second Edition by Ethem Alpaydın.
7. https://stats.stackexchange.com/questions/253926/what-are-the-x-and-y-axes-of-clustering-plots
8. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics, 28(1), 100. doi:10.2307/2346830
9. Rish. An Empirical study of the naive Bayes classifier. T.J. Watson Research Center
10. S.L. Ting, W.H. Ip, Albert H.C. Tsang. Is Naïve Bayes a Good Classifier for Document Classification?
11. Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. ACM SIGMOD Record, 36(2), 7–12. doi:10.1145/1328854.1328855
12. Akiko Aizawa,An information-theoretic perspective of tf–idf measures,Information Processing & Management,Volume 39, Issue 1,2003,Pages 45-65,ISSN 0306-4573,https://doi.org/10.1016/S0306-4573(02)00021-3.
13. T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967. doi: 10.1109/TIT.1967.1053964
14. Machine Learning by Tom M. Mitchell.
15. Niwattanakul, Suphakit et al. “Using of Jaccard Coefficient for Keywords Similarity.”
16. Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning - ICML ’06. doi:10.1145/1143844.1143874
17. Goutte C., Gaussier E. (2005) A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: Losada D.E., Fernández-Luna J.M. (eds) Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science, vol 3408. Springer, Berlin, Heidelberg

18. M. M. L. Elahi, R. Yasir, M. A. Syrus, M. S. Q. Z. Nine, I. Hossain and N. Ahmed, "Computer vision based road traffic accident and anomaly detection in the context of Bangladesh," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, 2014, pp. 1-6. doi: 10.1109/ICIEV.2014.6850780
19. Maniruzzaman, Khandoker & Mitra, Raktim. (2005). ROAD ACCIDENTS IN BANGLADESH. IATSS Research. 29. 10.1016/S0386-1112(14)60136-9.
20. M. S. Satu, S. Ahamed, F. Hossain, T. Akter and D. M. Farid, "Mining traffic accident data of N5 national highway in Bangladesh employing decision trees," *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dhaka, 2017, pp. 722-725. doi: 10.1109/R10-HTC.2017.8289059
21. M. M. L. Elahi, R. Yasir, M. A. Syrus, M. S. Q. Z. Nine, I. Hossain and N. Ahmed, "Computer vision based road traffic accident and anomaly detection in the context of Bangladesh," *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, Dhaka, 2014, pp. 1-6. doi: 10.1109/ICIEV.2014.6850780
22. Maniruzzaman, Khandoker & Mitra, Raktim. (2005). ROAD ACCIDENTS IN BANGLADESH. IATSS Research. 29. 10.1016/S0386-1112(14)60136-9.
23. Ahmed, Ishtiaque & Ahmed, Bayes. (2013). Urban Road Accidents in Dhaka, Bangladesh. 10.13140/2.1.2199.9683.
24. Rajaraman, A.; Ullman, J. D. (2011). "Data Mining". Mining of Massive Datasets (PDF).*Pp.*doi:10.1017/CBO9781139058452.002. ISBN 9781139058452.
25. www.wikipedia.org
26. The Nature of Code by Daniet Shiffman
27. https://dl.howtocode.com.bd/
28. Yujun Yang, Jianping Li and Yimei Yang, "The research of the fast SVM classifier method," 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, 2015, pp. 121-124. doi: 10.1109/ICCWAMTIP.2015.7493959
29. V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.
30. http://data.gov.bd/dataset/road-accident-and-casualties-statistics-2009-2016/resource/a63cf2e4-137b-42b3-9780
31. https://www.bbc.com/news/world-asia-45097650
32. https://www.thedailystar.net/ : July 02,2017
33. https://www.who.int/violence_injury_prevention/road_safety_status/2015/Road_Safety_SEAR_3_for_web.pdf