# Collecting traffic accident data by scrapping online national newspapers

**Mir Abir Hossain[1], Dr. Khondaker A. Mamun[2]**

*[2]Director of Advanced Intelligent Multidisciplinary Systems Lab, Professor, United International University, Dhaka, Bangladesh*

## 1. Abstraction

Traffic Accident is a big concern in Bangladesh. Every year many people die and get injured in traffic accidents and Bangladesh carries a huge loss. To eradicate traffic accidents, the first need is to collect traffic accident data properly. By analyzing those data, accident researchers will be able to figure out the reasons for accidents and how to eradicate them. But the current accident data collection system is erroneous and not all accidents get reported. An alternative solution to the traffic accident data collection process is to collect the data from newspapers. As it is highly unlikely that an accident happened and no newspaper has reported it at all. In this project, we will use web scrapping to collect accident news from online versions of national newspapers and create a database of traffic accidents that happened during a particular period of time.

## 2. Introduction

- Traffic accident is the leading cause of death for the age group of **15** to **29** in Bangladesh
- According to BUET ARI on average **12000** death per year and **35000** injuries per year occurs [2]
- Estimated total cost of accident is **4,118 million USD** ( **35000 crore BDT** ) which constitutes **1.3 %** of the total Gross Domestic Products (GDP) of Bangladesh.

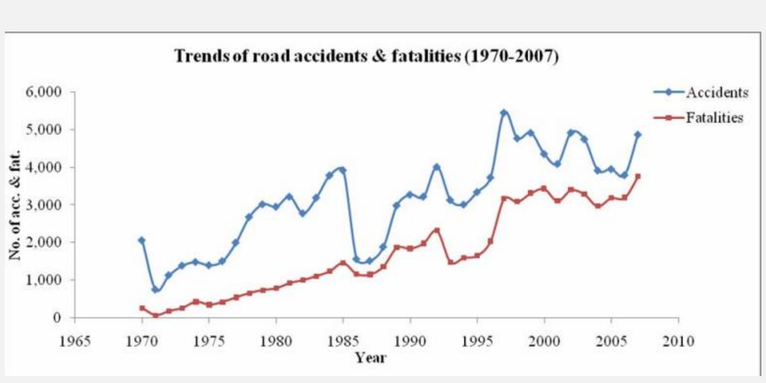


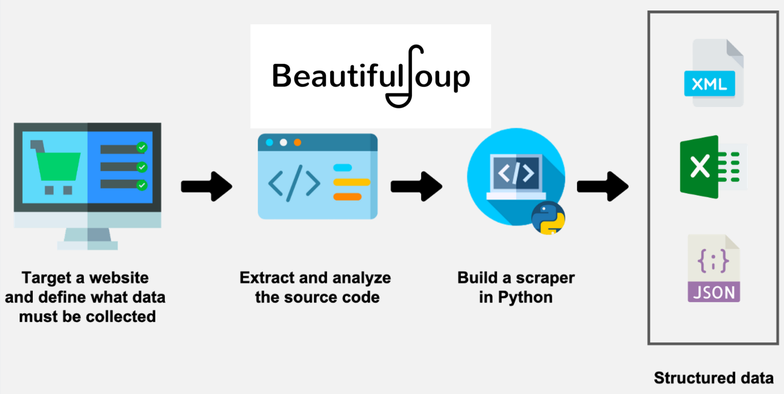

Figure: Police reported Trend of Road Accidents & Fatalities (1970-2007)

## 5. Web Scrapping

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. There are several tools to web scraping.

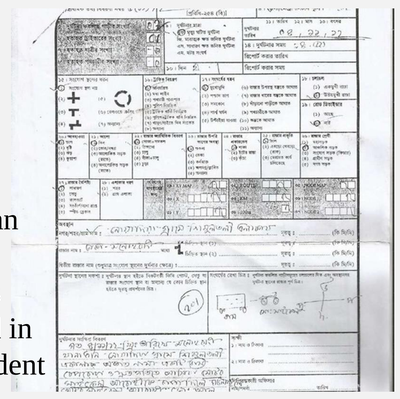
- Octoparse
- Parsehub
- Import.io
- BeauitifulSoup (Python library)
- Selenium (Python library)

We have used BeautifulSoup python library. It is flexible and is suitable for scraping data with poor website designs, and websites who changes their structure frequently.



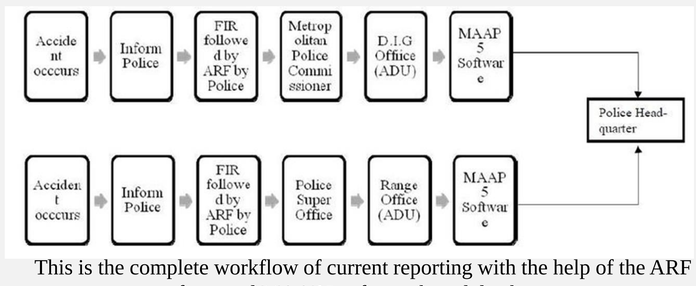

## 3. Current Reporting

Police use this ARF form for reporting an accident



Limitations:

- Can not represent actual accident situation fully
- Not all kinds of accident can be described perfectly
- Chances of mistake, as it is handwritten and performed in location at the time of accident discovery
- This forms can get lost anytime.




This is the complete workflow of current reporting with the help of the ARF form and MAAP5 software based database

## 7. Proposed Solution

As newspapers keep track of every accident diligently and properly, and there are quite a number of newspapers, it is a great source of authentic traffic accident data. Thus scrapping information from newspapers will create a more robust traffic accident database.

In our workflow we will use:

- Python
- BeautifulSoup
- Named Entity Recognition

Scrap news from newspaper ( By writing codes in python using the python library beautifulsoup) and save them in a csv file.

Filter news using specific keywords to get only traffic accident related data such as “traffic accident” “road accident” etc. Delete the rest of the news.

Using named entity recognition to get the accident type, injured count, death count, vehicle type and other important details from every news and save them in a csv file.

## 8. Next Step

We have only scraped one newspaper. Which has some limitations. Some accident may not have reported in one specific newspaper. To eradicate this limitation, we can use multiple newspaper.

Scrap multiple newspaper and filter news

Do NER for all filtered news

Check news at same date and same location and same vehicles. Then delete the duplicate news.

Removing the duplicate news would be a difficult task and need to perform high level NLP. This can be kept as future work.

## 4. Current Challenges

Current process is subjected to **Under-reporting.** There is a gap between actual traffic accidents, and reported traffic accidents. The following portion is taken from BRTA annual report 2005.

**2.3 INTERPRETATION OF DATA**

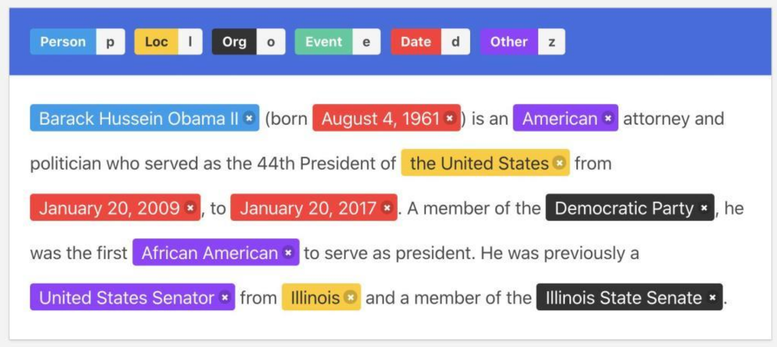
For targeting road safety improvement initiatives, interpretation of the accident data presented herein to either establish accident profiles or compare accident rates by District, Division or City can be undertaken with a measured degree of confidence. Caution is advised however when making comparisons of safety performance with that of other countries or when endeavouring to determine an absolute value of total accident occurrence.

BRTA and the Road Safety Cell are aware of a possibly considerable level of under-reporting of road traffic statistics: not all traffic accidents are registered on a “First Information Report”, and not all accidents reported as FIRs are entered on an Accident Reporting Form.

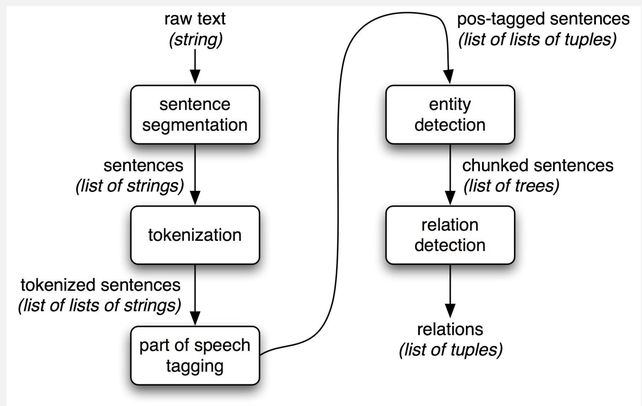
This problem does not have any easy solution. Making the police force stronger or launching new force with better capabilities is not an time consuming endeavor. Which makes classical solution of this problem almost impossible.

## 6. Named Entity Recognition NER

Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc It is an a Natural Language Processing (NLP) technique.




The workflow works like this




## 9. References

1. Saleh, Shameer. (2014). Identification of Hazardous Road Locations Using Different Methods on National Highways of Bangladesh.
2. T. Rahman, "Road Accidents in Bangladesh: An Alarming Issue", The World Bank, 2012.
3. 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.
4. Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.
5. John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001

*[1]MSCSE Student, United International University, contact email: mhossain222006@mscse.uiu.ac.bd*