**A Project Based Seminar Report**

**on**

**“Crime Analysis using newspaper articles”**

Submitted to the

Savitribai Phule Pune University

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

**Fatehraj Jetharam**

(Exam Seat No.:71825850K / Roll No.:42 & Division:T)

Under the guidance of

**Mrs. P .Y. Pawar**

****

Department Of Information Technology

Sinhgad Academy Of Engineering

Kondhwa,Pune , Maharashtra, India

**2020-21**

****

**CERTIFICATE**

This is to certify that the project based seminar report entitled “crime analysis using newspaper articles” being submitted by **Fatehraj Suthar(71825850K,42,T)** is a record of bonafide work carried out by him under the supervision and guidance of **Mrs. P.Y. Pawar** in partial fulfillment of the requirement for **TE (Information Technology Engineering) – 2015 course** of Savitribai Phule Pune University, Pune in the academic year 2017-2018

Date:

Place:

Mrs. P.Y. Pawar Prof. Abhay Adapanawar

Guide Head of the Department

This project based seminar report has been examined by us as per the Savirtibai Phule Pune University, Pune, requirements at Sinhgad Academy of Engineering on . . . . . . . . . . . . . .

(Name & Signature) (Name & Signature)

Internal Examiner External Examiner

I

**ACKNOWLEDGEMENT**

I am really grateful and wish my profound indebtedness to **Mrs. P.Y. Pawar,Professor,** Department of Information Technology,Sinhgad Academy of Enfineering, Pune.

Deep Knowledge & keen interest of my guide in the field of “*machine learning*” has helped me to successfully complete this project. Her endless patience ,scholarly guidance ,continual encouragement, constant and energetic supervision, constructive criticism , valuable advice and correcting my mistakes at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Prof. Abhay Adapanawar** , Head of Department of Information Technology, for giving me an opportunity to carry out the research work, without him I would’t have reached my goal and also to other faculty member and the staff of IT department of Sinhgad Academy of Engineering.

I would like to thank my entire course mates in Sinhgad Academy of Engineering, who took part in this discussion while completing the course work.

Finally,I must acknowledge with due respect the constant support and patience of my parents.

(Students Name & Signature)

II

**Abstract**

It is always a good choice to consider multiple sources when analyzing crime. Newspaper can be a good source of crime analysis as it always contains the massive amount of data related to any particular subject but the fact is they are not structured enough to make a definite decision.

In this paper, we represent a system to analyze crime news from online newspaper using machine learning techniques. These techniques help to extract useful pieces of information about crime from unstructured crime news .The analysis will provide the definite information about a news whether it is a crime related news or not, crime location, similarity and grouping crime news as stories based on a crime news.

The data is collected from newspaper’s online version and saved in the database .

Preprocessing is done and data is send to SVM classifiers which classifies news as crime or non-crime.

If the news is classified as crime. Then, crime location from that article is extracted and the article heading is collected for future references.

III

**LIST OF FIGURES**

**Sr. No. Figure Name Page No.**

1 Flowchart 12

2 news document classification 16

3 crime location extraction 16

4 ranking location names by occurrence of crime 17

IV

**Contents**

Certificate I

Acknowledgement II

Abstract III

List of Figures IV

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | **Sr.** | | **Chapter** | **Page No** | | **1.** |  | **Introduction** |  | |  | 1.1 | Introduction to Project | 7 | |  | 1.2 | Motivation behind Crime Management | 8 | |  | 1.3 | Objective of the work | 9 | |  | 1.4  1.5 | Introduction to crime analysis using newspaper articles  1.4.1 Aim and Objective of crime analysis using newspaper articles  Organization of the report | 9 | |  |  |  | 10 | | **2.** |  | **Literature Survey of crime analysis using newspaper articles** |  | |  | 2.1 | Introduction | 13 | |  | 2.2 | Different Approaches | 14 | |  | 2.3 | Algorithm | 15 | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | | **3** |  | **Applications of crime analysis using newspaper articles** | 16 | |  | 3.1 | State of Art applications |  | |  | 3.2 | Advantages and Disadvantages |  | | **4.** |  | **Conclusion** | 17 | | **5** |  | **References** | 18 | |  |  |  |  | |  |  | Crime News Analysis:Location and Story Detection | 19 | |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**CHAPTER 1**

**INTRODUCTION TO** **CRIME MANAGEMENT**

* 1. Introduction to Project

Crime Management is a complex task as it deals with a lot of unstructured data and different complexities at each level.Effective, responsible crime prevention enhances the quality of life of all citizens.Well planned crime prevention strategies promote community safety and contributes to sustainable development of countries.

It is widely recognized that we live in an increasingly unsafe society, but the study of security and crime prevention has lagged behind in its importance on the political agenda and has not matched the level of public concern.

This exciting new series aims to address these issues looking at topics such as crime control, policing, security, theft, crime and fear of crime.

With the advancement in criminal data, it has become difficult to maintain records manually and there is a need of computerized management of crime records, it has become a necessity to come up with modern solutions to tackle it .

The **crime control model** focuses on having an efficient system, with the most important function being to suppress and control crime to ensure that society is safe and there is public order. Under this model, controlling crime is more important to individual freedom.

Machine Learning is one such tool that can assist law enforcement bodies to prevent and predict various types of crimes. but with proper training and correct algorithms, it can be simplified and can become a great tool in fighting crime.

 We can identify the root causes of crime and make earnest efforts for crime prevention.

We have tried to implement a system that seeks to reduce the risk of occurrence of crime and their potential harmful effects on individuals and society.

* 1. Motivation behind project topic

With time, crime is growing like a cancer and corrupting every place it sees.

There is an increasing demand for the application of crime analysis techniques that utilize data available from new papers and create patterns in crime occurrence to assist in crime management.

There is Revolution of Data Analysis along with Big Data, which is changing the phase of computing.

Machine learning is a tool that can ease lot of manual work and there is lot of opportunity to learn new things.

Crime management reflects the nation’s condition globally and its proper management can save the nation from various horrors.

Even though we are in 21st century but still we as a nation lack advanced techniques to fight crime and this is the sole reason for choosing this topic .

* 1. Aim and Objective(s) of the work

Project aim :

The aim of this project is to understand working of various machine learning algorithms and their applications that can be used to implement a crime management system to aid law enforcement bodies in analyzing and predicting crimes.

Project objectives :

* To study crime based dataset.
* To Develop and seek Knowledge on how to deal with Online Data Source.
* To learn to select the suitable features for training of the models.
* To use machine learning algorithms to analyse and generate patterns in crime.
* Try to identify the most Crime Affected spot in a certain region.
* Try to predict the crime attempt in a certain area.
* Try to reduce crime and disorder.

* 1. Introduction to Crime Analysis using newspaper articles

1.4.1 Aim and Objectives of Crime Analysis using newspaper articles

* The aim is to extract information regarding crime from various newspaper articles and classify it using 2 class classifier.
* Objectives :
  1. To select enlightening features related to crime.
  2. To collect news related data from newspaper’s online version.
  3. To train and test the SVM classifier model using the selected features.
  4. To classify news as crime or non-crime.
  5. To extraction crime location from news document.
  6. Rank location names by the occurrence of number of crimes.
  7. Show similarity based on a given article.
  8. To provide aid to the law enforcement bodies in analyzing crime.
  9. Organization of the report

The prime goal of this was to create a system which can classify news from news articles.

WORKING

1. DATA COLLECTION:
2. collect news from newspaper’s online version using web crawler.

b) Save data in the database:

MongoDB is a NoSQL database that provides data model convenient to design and promises high performance and scalability. So it is the best-suited database program for our system. We have used driver for MongoDB to handle database operations.

c) Data preprocessing:

- Clears the structure of text data.

- Word tokenization, removing stop words and special characters, case conversion, lemmatization is performed on each document before saving in database for processing purpose(NLTK library).

- NLTK is natural language toolkit that can be used to process English grammar words.

- Word tokenization is performed before case conversion. A word matching process is performed for each news text and words not in the dictionary are collected. Then few words that have no effect are listed. We have used NLTK library’s builtin English stop words combined with our stopword list and removed those from each news text. This is a temporary state not permanently saved. Lemmatization is performed on these text data.

- A collection of pre-processed terms is kept in the database.

-Feature selection and feature transformation improves the performance of the classification process.

- Each term with their term frequency for a document and their total document frequency is calculated and saved to make the next text analysis steps faster and easier.

-Feature selection techniques TF-IDF is used to select the features that are highly discriminative.

1. NEWS DOCUMENT CLASSIFIER:

a) The proposed document based SVM classification system takes the documents and generates a model that classifies the new documents as crime or non-crime.

b) I have tried both linear and rbf kernel to compare the performance.SVM can give better performance when hyper-parameter is tuned properly.

c) I have used grid search with 5-fold cross-validation.

d) In k fold cross validation, iteratively , each fold is used as a testing set at some point and the rest of folds are used to train the model until each of the 5 folds have been used as the testing set.

1. LOCATION NAMED EXTRACTION:

a) Crime location extraction is performed on documents using named entity recognition (NER) and sentence classification.

b) Named Entity Recognition: Sentence tokenization is performed to get all the sentence from a document. For each line word tokenization and part-of-speech tagging is performed. Then NER is performed over each sentence to get the sentence with location names. The named entity recognition is performed using two different ways. We have made a list by collecting location names (city, district, division) from our downloaded newspaper dataset and wikipedia. The combined list used as a custom list.

c) A sentence classifier is implemented to classify the sentence which has probable crime location. Sentence with a location in crime news is used as training data.

d) A grid search is performed to tune the hyper-parameters before classification.

e) Crime Location Ranking: Top location names by the number of crime occurred are ranked and listed. The crime location ranking is implemented to perform the operation on a collection of crime news in the database and output the rank in sorted order.

1. GROUPING NEWS STORIES:
2. Searching Similarity: The news similarity in our local database is implemented based on cosine similarity.
3. The query document is converted into tf-idf vector then normalized. The

dot product between each document in database and query document is used as a measure of ranking similarity between query and database documents.

b) As all documents at least have few obvious minor similarities among them we introduced a threshold value of cosine similarity to keep ranking until that value.

c) Clustering News Story: To group same type of story based one a query news document, we have used clustering algorithms after similarity ranking. The hierarchical clustering is implemented using cosine similarity.

1. RESULT:
   * 1. The data from news sites is collected by the crawler.
     2. Few garbage texts (not news document) are collected by the crawler
     3. The classifier accuracy evaluation for same number of news documents for two different newspapers is found.

Crime location name extraction is done for a single document and top crime location names ranked by the number of crime occurred in that location are found.

* + 1. An article is searched and similarity to the searched document string from the database is found.
    2. List of a very similar documents from the same crime story clustered by the system and ordered by the published date are found. Each news following the previous news. Range of date in which the related news of the group published and crime location list extracted from each news in the group for the only cluster that contains the news closest to the searched news.

F. OUTPUT: a) NEWS document classification

b) Crime location extraction from news document.

c) Ranking location names by the occurrence of number of crime.

d) Ranking of crime news by similarity.

e) Searched news story (cluster).

In short, to summarize :

* First we collect data using web crawler from different news sites as it facilitates real time analysis of news data.
* Collected data is saved in the database using MongoDb
* Data is preprocessed to reduce the size of file
* Lemmatization is produced using NLTK library to remove stop words
* Case conversion is done after this
* Feature selection
* Feature transformation
* We use k-fold cross validation for training and testing data set.
* SVM classifier gets the processed data and classifies crime
* If the classifier classifies the article as crime related ,the location extraction is done using Named Entity Recognition to store the location where the crime took place
* Then,we classify that sentence using sentence classification.

**CHAPTER 2**

**LITERATURE SURVEY OF CRIME ANALYSIS USING NEWSPAPER ARTICLES**

2.1 Introduction

This section provides an overview of the related work in the domain of crime extraction.

Crime monitoring and prevention is a domain of interest to all countries across the globe in order to make the world a safe place to live. Use of ICT technologies for this purpose has been around since the advent of computers. Currently, with wide spread availability of news articles (e.g. through freely available online newspapers), re-searchers have become interested in combining these information to monitor and prevent crimes.

Researchers working in the area of crime information extraction have used several techniques. In particular, researchers have used techniques such as crowd sourcing, data mining and machine learning for this purpose.

Researchers have explored techniques for retrieving relevant information from unstructured documents .The process of extracting information from unstructured documents is difficult because it is written in natural language and the structure of the document is not known ahead of time (when compared to structured \_les such as databases). However, there has been a lot of work on identifying entities (e.g. person, place, organization) from unstructured documents in the field of natural language processing.

It is always a good choice to consider multiple sources when analyzing crime.

The newspaper is a vast information source but the information is not organized or specific to take decisions based on them. Crime analysis from the newspaper article is an area of processing and refining to extract unknown information.

Detecting crime news, crime location, grouping similar crime news are few of them which can be used as basic methodologies for extracting information to detect crime pattern and criminal behaviour.

The analysis will provide the definite information about a news whether it is a crime related news or not, crime location, similarity and grouping crime news as stories based on a crime news.

SVM classification algorithm is used to perform a two-class classification when collecting data from online newspapers.

The common techniques like word tokenization, removing stop words and special characters, case conversion, lemmatization performed on each document before saving in database for processing purpose. All of these operation is performed using NLTK library.

2.2 Different Approaches :

1. Roja bandari et al. analyze the news articles and predict the online popularity of that articles using support vector machine.
2. Steven Pires and Ronald Belance discusses the crime problems faced by the police departments in and around the college campus. Geographic Information System is used to identify the spots for criminal behavior has been mostly useful for police to reduce problem.
3. Monica Rogati and Yiming Yang performed an wide study of the performance of over 100 variants of 5 filter feature selection methods using two benchmark collections (Reuters 21578 and part of RCV1) and four classifiers (Naive Bayes, Rocchio, K-Nearest Neighbor and Support Vector Machines). The result of the study shows that the methods which include chi-square attain high performance in classification.
4. The Wiki Crimes project (Wiki Crimes 2013, Furtado et al. 2010). harnesses the power of the crowd, where individuals report crime details online and other users can use this information to make decisions. However, a limitation of this approach is the difficulty of verifying the authenticity of the posted crimes.
5. Tehseen Zia, Qaiser Abbas and Muhammad Pervez Akhtar conducted a study to analyze the performance of five feature selection method such as information gain, gain ratio, Chi statistics, symmetric uncertain and OneR using six classifiers (naive Bayes, KNN, support vector machine with linear, polynomial and radial basis kernels and decision tree) on two Urdu test collections: naive collection and EMILLE collection. It shows that Linear SVM with feature selection methods IG and Chi has outperformed other combinations of classifiers.
6. The Coplink project (Chen et al. 2004) of re-searchers at the University of Arizona aims at identifying information about criminals from police reports.It uses an Entity Extractor system that is based on AI techniques, for detecting identities of criminals automatically and also for analyzing criminal networks using clustering and block modeling.
7. The work of De Bruin et al. (2006) used an approach for analyzing criminal careers. Based on the analysis, they have identified four important factors (crime nature, frequency, duration and severity).

2.3 Algorithm

Steps for document classification :

1. Data collection.
2. Data preprocessing : to clean and reduce the data.
3. Construction of term-document matrix.
4. Feature selection.
5. Feature transformation.
6. Classification:
   1. If non-crime, the article is excluded.
   2. If it is crime
      * Extract crime location using Named Entity recognition.
      * Collect sentences that contain crime locations.
      * Sentence classifier is used to classify sentences and these sentences are used in training data.

Example:

|  |  |
| --- | --- |
| Sentence Type | Class |
| The sentence contains location and previous sentence contains crime related terms | Crime Location |
| The sentence contains crime related terms with location | Crime Location |
| The sentence contains victim’s or criminal’s home address (city,town names) | Not Crime Location |
| The sentence contains police station location | Crime Location |
| The sentence contains the hospital name or the location where victims have been taken | Not Crime Location |

* + - Crime location ranking based on the occurences of crime in that area.
    - Grouping news stories.

**CHAPTER 3**

**Applications of crime analysis using newspaper articles**

* 1. Applications

* The system can be used to show regional crime trends.
  + We can port these subsystems to any crime analysis system to perform a basic analysis.
  + Crime news detection, crime location extraction and story detection
  + To understand crime pattern in a single location.
  + Can be used to understand the density of a single type of crime.
    1. Advantages
* Ease in crime related news extraction
* Gives better accuracy as compared to naïve bayes models
* Works well with unstructured and semi structured data
* It scales relatively well to high dimensional data.
* SVM models have generalization in practice, the risk of over-fitting is less in SVM.  
  + 1. Disadvantages
* 1st and most drawback is difficulty in obtaining data
* System takes long training time for large datasets
* When determining the threshold of cosine similarity

ranking, we have used an average of first dissimilarity

point but this may differ in the situation where a dataset

which only contains very similar news documents of same

stories and very dissimilar type of stories (named entities and crime types).

* Training data set is collected manually for document

classification and named entities (location, human) are

removed manually. It is very much possible for very

larger training data to have some entities that can affect

the result of classification though we have ignored those terms that has small number of document occurrence.

**CHAPTER 4**

**CONCLUSION**

Looking back we have successfully classified news document using 2 class classifier.

We collected data using crawler and then saved it to database.Then,data preprocessing is done and transformed data is passed as input to SVM classifier which classifies the news as crime or non-crime.We sort crime location as per crime occurences.

So,we conclude that:

* SVM classifier correctly classifies the news as crime / non-crime
* The system provides different features to analyze crime through the online newspaper.
* Grouping stories show similar crime stories and linked news for the crime story to be concerned. It helps to understand the density of a single type of crime.

**REFERENCES**

List all the material used from various sources for making this project proposals

1. Crime News Analysis:Location and Story Detection , 2017 IEEE

* Link: <https://ieeexplore.ieee.org/document/8281798>

1. News document analysis by using a proficient algorithm , International Journal of Engineering Research and Applications (IJERA) *ISSN : 2248-9622, Vol. 7, Issue 6, (Part -2) June 2017*

* Link: [http://www.ijera.com/papers/Vol7\_issue6/](https://ieeexplore.ieee.org/document/7112359)

1. Crime Analytics: Analysis of Crimes Through Newspaper Articles,2015 IEEE
   * Link: <https://ieeexplore.ieee.org/document/7112359>
2. Crime Analysis against Women from OnlineNewspaper Reports,2017 IEEE
   * Link: <https://ieeexplore.ieee.org/document/8070855>

978-1-5386-1150-0/17/$31.00 ©2017 IEEE

BASE PAPER

Crime News Analysis:

Location and Story Detection

|  |  |
| --- | --- |
| Mehedee Hassan  *Department of Computer Science and Engineering*  *Jahangirnagar University*  Dhaka, Bangladesh mehedee.hassan@outlook.com | Mohammad Zahidur Rahman  *Department of Computer Science and Engineering*  *Jahangirnagar University*  Dhaka, Bangladesh rmzahid@juniv.edu |

*Abstract*—It is always a good choice to consider multiple sources when analyzing crime. Newspaper can be a good source of crime analysis as it always contains the massive amount of data related to any particular subject but the fact is they are not structured enough to make a definite decision. In this paper, we represent a system to analyze crime news from online newspaper using different data mining techniques. These techniques help to extract useful pieces of information about crime from unstructured crime news. The analysis will provide the definite information about a news whether it is a crime related news or not, crime location, similarity and grouping crime news as stories based on a crime news.

*Index Terms*—information extraction, crime analysis, machine learning

I. INTRODUCTION

The way crime is increasing so as the crime related data it’s becoming harder to analysis crime information day by day. The ever-growing online newspaper archives added a new dimension to crime data analysis. The newspaper is a vast information source but the information is not organized or specific to take decisions based on them. Different types of difficulties are introduced when processing these unstructured data. So this crucial source of information often remains unutilized. Common data mining techniques can be used to solve these problems. Crime analysis from the newspaper article is an area of processing and refining to extract unknown information. Detecting crime news, crime location, grouping similar crime news are few of them which can be used as basic methodologies for extracting information to detect crime pattern and criminal behavior.

Not all online newspaper’s news is categorized as a crime news. To analyze crime, at first, we need to collect crime news. So classification algorithm is used to perform a two-class classification when collecting data from online newspapers. It will be a time consuming and irritating procedure to collect crime news without an automated system.

When analyzing crime news the crime location is important information. It can be extracted manually but will be error prone task when a large number of documents to be processed. To reduce the complexity we have used sentence classification and named entity recognition. As a news may contain location names in more than one lines a classification model to classify only sentences that contain location name related to crime will

be helpful [1]. It is an awful job to find similar news from a large set of news. Search engines can be used to find similar news but it will only provide straight news articles ranked related to the search string. The problem is, there will be few articles where the main stories are not related to the searched one. Moreover, there will be few news articles which have a stronger relation to one another than the searched one.

Considering the above situation we have implemented a way to solve this problem by finding similarities based on a document and then clustering based one similarity among the ranked document to group them into different stories where they have more similarities among the news articles inside a

group. To be specific, the story of a set of news is extracted based on a specific crime news.

The paper is organized as follows. A short description of related works is in Section II. Definition and short description of key concepts to understand the system is discussed in Section III. The summary of methodologies used to implement the system is in Section IV. The experiment with the real world environment and their results are shown in Section V. Weaknesses of the system are pointed in Section VI. The conclusion and future steps discussed in Section VII.

II. RELATED WORKS

Understanding the crime is the initial part of crime analysis. An efficient crime analysis performed in [2], they discussed several types of crimes. These crimes can be differentiated into two levels local law enforcement level, national and international law enforcement level. Though very few types of crime may have an impact on both national and international level. Few areas like immigration and transportation heavily depend on crime analysis as we avoid crime prone areas [1]. Crime analysis can be performed using various data mining techniques [2], [3]. Few common and effective techniques are classification, entity extraction, clustering. Few more techniques are used like association rule mining deviation detection and sequential pattern mining [3].

A criminal network analysis is to extract criminal relations from the incident summaries and create a likely network of suspects is implemented in [2]. They also performed true identity detection of a criminal with the help of police database. In [3], they implemented crime pattern visualization, crime comparison graph and crime hotspot analysis. Another inspiring work [4], implemented step by step data mining techniques like classification, named entity extraction to identify crime patterns. Then crime hot spots with crime types are also implemented. They used decision tree to predict crime. A heat map is used to visualize crime prone areas. In [5], clustering techniques used to locate the crime clusters in the area of interest. Later it helps to detect crime hot spots and crime zones of a particular region on certain crime types for the period specified by the user. In [6], three different clustering algorithms with other common data mining techniques are discussed. They discussed few challenges and research gaps that are still to be fulfilled in crime analysis process.

In [7], they introduced a distance measure to incorporate the frequency of delinquent behavior into the matrix which is then summed over the available years. The comparison to calculate distance is based on profile, crime severity and the number of crimes and clustered using a human-centered multidimensional clustering approach to determine the criminal career. Crime information extraction from an online newspaper can be complex and error prone if performed manually. It can be performed using step by step processes. These techniques are shown in [1] and [3]. In [1], they started from building corpus then feature identification and extraction performed for classification. They also performed location extraction using named entity recognition libraries.

All the works we have discussed above introduced several techniques about crime news classification and location extraction. They didn’t try to extract the crime story. In our system, we have introduced a feature to extract a set of follow up news based on a specific given crime news.

III. DEFINITION OF TERMS

# A. Document Classification

Document classification that is plain text document classification literally means assigning text document into one or more classes (text document categories) assuming that we have a set of predefined classes [8]. A lot of algorithms are suitable to use in text document classification. SVM is preferred on other classifiers because it can achieve good performance even if text classification problems involve in high dimensional feature spaces [9]. When considering each word in the training text document as a feature, training document of a smaller amount may have a lot of attributes [9]. Text documents are preprocessed before classification that provides well-defined sequences of linguistic components that have standard structure and notation [8].

# B. Named Entity Recognition

Named Entity Recognition refers to recognizing the small phrase from text [10]. A text segment may have few terms represented as entities which are more informative and unique in contextual purpose called named entities [8].Common named entities are a person (Hasan), an organization (WHO), location (Bangladesh), date (June,2017-06-01), time (12:00 am) and so on. Named entity recognition is literally the process of extracting the named entity and classifying them to a predefined class. It is also known as entity chunking/extraction which is a part of information extraction [8], [11].

# C. Document Clustering

Document clustering is an unsupervised process of grouping documents into different groups called cluster [11]. The uncategorized documents are grouped into the meaningful clusters without any prior information. There are many clustering algorithms that can be implemented to cluster documents. Hierarchical clustering is a common algorithm that is used for text document clustering. It always decomposes the set of data to be clustered [12].

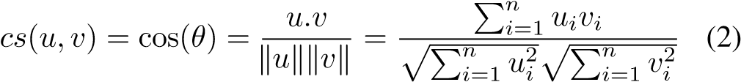
The general concept is to successively merge two clusters depending on their similarity. After each step total number of cluster decreases. In each step, we may have different size of clusters. A single document considered as the smallest sized cluster. The process (merging two clusters) is continued until the number of clusters becomes one.

# D. Similarity Measure

Similarity measures are used for text similarity analysis which measures the degree of closeness between two text bodies. It can be useful in identifying similar entities and distinguishing clearly different entities from each other. Satisfying few conditions. A distance measure of similarity can be called a distance metric of similarity [8]. A common and easy to understand measure of similarity is cosine similarity.

For two document’s terms represented in their vectorized forms, cosine similarity gives us the measure of the cosine of the angle between them when they are represented as nonzero positive vectors in an inner product space. The cosine similarity (cs) of two-term vector u,v can be defined as the dot product of the two term vectors u and v, divided by the product of their L2 norms. Mathematically represented as,

*u.v* =k*u*kk*v*kcos(*θ*) (1)



Where, *θ* is the angle between u and v .k*u*k represents the L2 norm for vector u and k*v*k is the L2 norm for vector v. Let say p and q is respectively u divided by k*u*k and v divided by k*v*k then we can write,

*cs*(*u,v*) = cos(*θ*) = *p.q* (3)

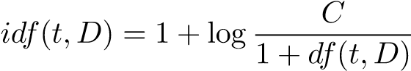
Equation (3) shows, if two term vectors are divided by their L2 norm then the cosine distance is simply the resulted vectors dot product.

# E. Feature Selection and Extraction

Features are unique measurable attributes or properties for each observation in a dataset. Feature selection is a process of selecting important features as a subset from the original feature set [13].

Extracted features are used in machine learning algorithm to find patterns in the data set. These algorithms usually perform operations on numeric vectors to learn patterns [8]. So numeric features vectors are expected as input. So, we have to transform textual data and extract numeric features from it. There is a common mathematical and algebraic way for transforming and representing text documents as numeric vectors of specific terms that form the vector dimensions called vector space model or term vector model. Text documents can be represented in vector space by making a vector of each distinct term’s weight.

One of the most common term-weighting schemes is TFIDF. The Term Frequency-Inverse Document Frequency (TFIDF) is a product of two metrics: term frequency (TF) and inverse document frequency (IDF). The TF in a document vector space is the term frequency value for a particular document (d). The IDF of a term (t) is computed by dividing the total number of documents (C) in a document set (D) by the document frequency (df) for each term and then applying logarithmic scaling on the result. An extra 1 is added to avoid total value zero. Mathematically IDF of term t is,

 (4)

Where, df(t,D) represents the document frequency for term t in a document set D. TF-IDF is calculated as, *tf* − *idf*(*t,d,D*) = *tf*(*t,d*)*.idf*(*t,D*) (5)

IV. METHODOLOGIES AND EVALUATION

# A. Data Collection

As we are working with online newspaper, we have used online English-language newspaper of Bangladesh for both training and testing purpose. We have collected news from

“The Daily Star” and “New Age” newspaper’s online version.

1. *Crawler:* We have implemented a crawler using simple tools and libraries like selenium and newspaper to collect plain text from online newspaper [14], [15]. The crawler only takes the title , the body contents and the date of a news. The crawler run within a specific domain to collect data. At a time it will only collect data from a single domain address.
2. *Saving Data:* We have used MongoDB as a database backend for saving text data collected by the crawler. As MongoDB is a NoSQL database that provides data model convenient to design and promises high performance and scalability [16]. So it is the best-suited database program for our system. We have used driver for MongoDB to handle database operations.
3. *Data Pre-processing:* Data pre-processing is performed to make a clear structure of text data. The common techniques like word tokenization, removing stop words and special characters, case conversion, lemmatization performed on each document before saving in database for processing purpose. All of these operation is performed using NLTK library [17].

Word tokenization is performed before case conversion. A word matching process is performed for each news text and words not in the dictionary are collected. Then few words that have no effect are listed. We have used NLTK library’s builtin English stop words combined with our stopword list and removed those from each news text. This is a temporary state not permanently saved. Lemmatization is performed on these text data.

A collection of pre-processed terms is kept in the database. Each term with their term frequency for a document and their total document frequency is calculated and saved to make the next text analysis steps faster and easier.

# B. News Document Classifier

To classify news collected from online newspaper we built a two-class classifier. The classifier classifies news article as crime-related news and non-crime related news. We have used support vector machine (SVM) classifier. We have tried both linear and radial kernel to compare the performance. SVM can give better performance when hyper-parameter is tuned properly [18]. We have used grid search with 5-fold cross-validation to find the best value for gamma and C. We have used Scikit-Learn library on this purpose [19]. 20% of total data set is used for validation purpose. Table I shows classifier’s performance evaluation comparison data.

TABLE I

DOCUMENT CLASSIFIER PERFORMANCE EVALUATION AND COMPARISON

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kernel | Precision | Recall | F-measure | Accuracy |
| SVM (rbf) | 0.962 | 0.968 | 0.964 | 0.96 |
| SVM (linear) | 0.955 | 0.965 | 0.954 | 0.95 |

# C. Location Name Extraction

Crime location extraction is performed on documents using named entity recognition (NER) and sentence classification. We have extracted location using NER methods and collected sentences that contain location names. Then a sentence classifier is used to classify sentence with crime locations.

1. *Named Entity Recognition:* Sentence tokenization is performed to get all the sentence from a document. For each line word tokenization and part-of-speech tagging is performed. Then NER is performed over each sentence to get the sentence with location names. The named entity recognition is performed using two different ways. We have made a list by collecting location names (city, district, division) from our downloaded newspaper dataset and wikipedia [20]. The combined list used as a custom list. Table II shows comparative validation of both processes.
2. *Sentence Classification:* A sentence classifier is implemented to classify the sentence which has probable crime location. Sentence with a location in crime news is used as training data. Table III shows the class for each type of sentence considered. We have used SVM for this purpose. A grid search is performed to tune the hyper-parameters before classification. Table IV shows the comparative data for two kernels used with SVM.

TABLE II

ACCURACY EVALUATION AND COMPARISON OF NER METHODS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kernel | Precision | Recall | F-measure | Accuracy |
| NLTK (pre-trained named entity chunker) | 0.83 | 0.80 | 0.81 | 0.96 |
| NLTK (Chunk parser with gazetteers) | 0.84 | 0.85 | 0.84 | 0.96 |

TABLE III

SENTENCE TYPES AND CLASS

|  |  |
| --- | --- |
| Sentence Type | Class |
| The sentence contains location and previous sentence contains crime related terms | Crime Location |
| The sentence contains crime related terms with location | Crime Location |
| The sentence contains victim’s or criminal’s home address (city,town names) | Not Crime Location |
| The sentence contains the hospital name or the location where victims have been taken | Not Crime Location |
| The sentence contains police station location | Crime Location |

TABLE IV

SENTENCE CLASSIFIER PERFORMANCE EVALUATION AND COMPARISON

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kernel | Precision | Recall | F-measure | Accuracy |
| SVM (rbf) | 0.955 | 0.656 | 0.778 | 0.721 |
| SVM (linear) | 0.955 | 0.653 | 0.775 | 0.718 |

1. *Crime Location Ranking:* Top location names by the number of crime occurred are ranked and listed. The crime location ranking is implemented to perform the operation on a collection of crime news in the database and output the rank in sorted order.

# D. Grouping News Stories

Searching similarity based on a news is performed at first then grouping similar news as a story based on a crime news is performed.

1. *Searching Similarity:* The news similarity in our local database is implemented based on cosine similarity. The query document is converted into tf-idf vector then normalized. The dot product between each document in database and query document is used as a measure of ranking similarity between query and database documents.

As all documents at least have few obvious minor similarities among them we introduced a threshold value of cosine similarity to keep ranking until that value. Table V shows the average similarity value for documents inside our database. We have taken the average of these two values.

1. *Clustering News Story:* To group same type of story based one a query news document, we have used clustering algorithms after similarity ranking. The hierarchical clustering is implemented using cosine similarity.Table V shows an average threshold used to maintain minimum cluster distance.

Table VI shows cluster quality with evaluation measures (purity) for a news document from our database. Where it shows purity for a same type of story and for a same type crime news that has no relationship among them as a part of the same story (location, the victim, criminal).

TABLE V

THRESHOLD VALUE

|  |  |
| --- | --- |
| Type of Limit | Average Cosine Similarity |
| Same Story (by dissimilarity starting point ) | 0.21 |
| Same Type of Crime | 0.072 |

TABLE VI

CLUSTERING QUALITY EVALUATION

|  |  |
| --- | --- |
| Type | Purity |
| Same Story | 0.51 |
| Same Type of Story | 0.82 |

V. EXPERIMENTS AND RESULTS

We have provided validation, evaluation and comparisons for each subsystem to evaluate their performance in Section IV. The experimental result for the system in real world environment is discussed in this section.

We executed the crawler for two different newspaper’s website for same amount of time. Few garbage texts (not news document) are collected by the crawler. The percentage of really usable news documents for each site is given in Table VII .

The classifier accuracy evaluation for same number of news documents for two different newspapers is give in Table VIII. Fig. 1 shows the output of news classifier when a single news document is provided as an input. The classifier could correctly detect the input document as a crime related news.

Fig. 2 shows the output of crime location name extraction for a single document. The system detected ‘Dhaka’ as a crime location and ignored other two location names ‘Rampura’ and ‘Hazipara’. Fig. 3 shows the top crime location names ranked by the number of crime occurred in that location (ignoring location names outside of Bangladesh) in descending order. The database contains one year crime news data from The Daily Star newspaper. Fig. 4 shows the output of similarity based on a given article “CID Team visits...”. It shows a ranking by similarity to the searched document string from

TABLE VII

CRAWLER PERFORMANCE

|  |  |
| --- | --- |
| Newspaper | News Document (%) |
| The Daily Star | 85.1 |
| New Age | 90.5 |

TABLE VIII

CLASSIFIER PERFORMANCE EVALUATION

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Website | Precision | Recall | F-measure | Accuracy |
| The Daily Star | 0.95 | 0.96 | 0.96 | 0.95 |
| New Age | 0.97 | 0.95 | 0.95 | 0.94 |

Fig. 1. News document classification.

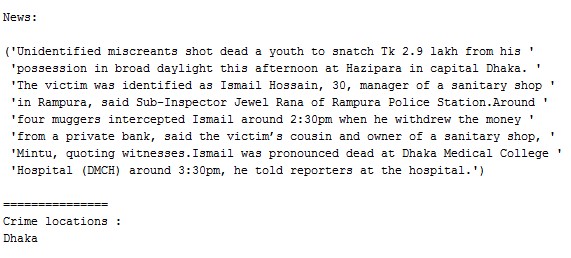
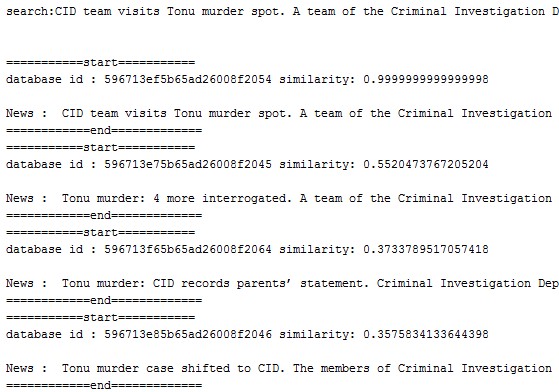
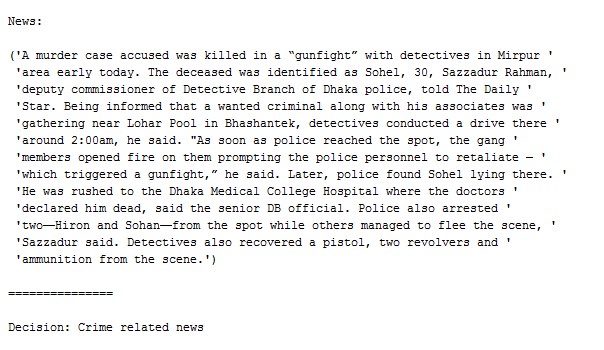


Fig. 2. Crime location extraction from news document.

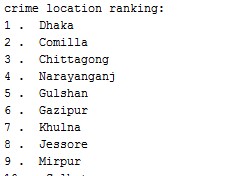


Fig. 3. Ranking location names by the occurrence of number of crime.

our database. We can see the first part of the output shows very similar news documents are ranked. The second part of the output shown in Fig. 5. It shows a list of a very similar documents from the same crime story clustered by the system and ordered by the published date. Each news following the previous news. It also shows the range of date in which the related news of the group published and the crime location list extracted from each news in the group for the only cluster that contains the news closest to the searched news. The operation is performed on crime news saved from The Daily Star newspaper of the year 2016 and 2017. The news story clustering quality for same number of documents from two

Fig. 4. Ranking of crime news by similarity.

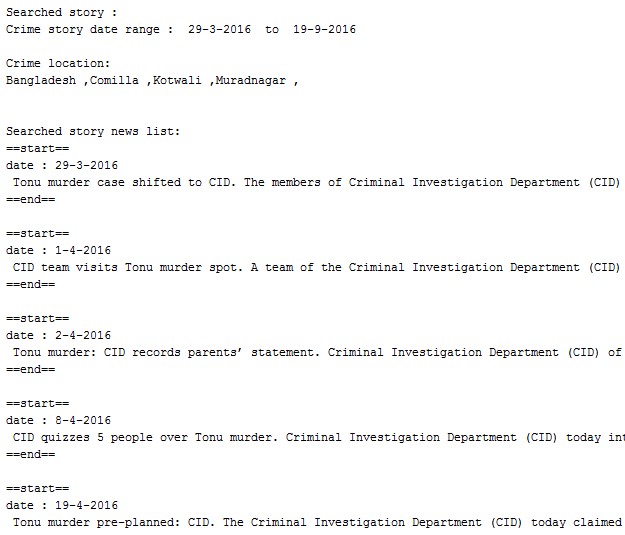


Fig. 5. Searched news story (cluster).

different newspaper is given in Table IX.

TABLE IX

STORY CLUSTERING QUALITY

|  |  |
| --- | --- |
| Site | Purity |
| The Daily Star | 0.48 |
| New Age | 0.50 |

VI. WEAKNESSES

The system certainly has few weaknesses and we are aware of that, they are ponied in below,

* When determining the threshold of cosine similarity ranking, we have used an average of first dissimilarity point but this may differ in the situation where a dataset which only contains very similar news documents of same stories and very dissimilar type of stories (named entities and crime types).
* Training data set is collected manually for document classification and named entities (location, human) are removed manually. It is very much possible for very larger training data to have some entities that can affect the result of classification though we have ignored those terms that has small number of document occurrence.

VII. CONCLUSION AND FUTURE WORKS

The system provides different features to analyze crime through the online newspaper. We can port these subsystems to any crime analysis system to perform a basic analysis. Crime news detection, crime location extraction and story detection may help to understand crime pattern in a single location. Ranking crime location shows the crime prone areas. Grouping stories show similar crime stories and linked news for the crime story to be concerned. It helps to understand the density of a single type of crime.

We will try to improve the efficiency of the system in future by overcoming the weaknesses. The system has a text-based user interface, the location ranking and the location detection is text based output. A map for top crime prone area will help more to understand visually. So we will try to add a map location and a graphical user interface. The crime type detection feature will be added as we already have training data of the common types of crime. Above all achieving more accuracy will be our main concern.

REFERENCES

1. R. Arulanandam, B. T. R. Savarimuthu, and M. A. Purvis, “Extracting crime information from online newspaper articles,” in *Proceedings of the Second Australasian Web Conference - Volume 155*, AWC ’14, (Darlinghurst, Australia, Australia), pp. 31–38, Australian Computer Society, Inc., 2014.
2. H. Chen, J. Jie, Y. Qin, and M. Chau, “Crime data mining: A general framework and some examples,” *IEEE Computer*, vol. 37, pp. 50–56, 2004.
3. I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, and A. Wijayasiri, “Crime analytics: Analysis of crimes through newspaper articles,” in *2015 Moratuwa Engineering Research Conference (MERCon)*, pp. 277–282, April 2015.
4. S. Sathyadevan, D. M. S, and S. G. S., “Crime analysis and prediction using data mining,” in *2014 First International Conference on Networks Soft Computing (ICNSC2014)*, pp. 406–412, Aug 2014.
5. B. Chandra, M. Gupta, and M. P. Gupta, “Adaptive query interface for mining crime data,” in *Proceedings of the 5th International Conference on Databases in Networked Information Systems*, DNIS’07, (Berlin, Heidelberg), pp. 285–296, Springer-Verlag, 2007.
6. U. Thongsatapornwatana, “A survey of data mining techniques for analyzing crime patterns,” in *2016 Second Asian Conference on Defence Technology (ACDT)*, pp. 123–128, Jan 2016.
7. J. S. D. Bruin, T. K. Cocx, W. A. Kosters, J. F. J. Laros, and J. N. Kok, “Data mining approaches to criminal career analysis,” in *Sixth International Conference on Data Mining (ICDM’06)*, pp. 171–177, Dec 2006.
8. D. Sarkar, *Text analytics with Python : a practical real-world approach to gaining actionable insights from your data*. United States New York, NY: Apress,Distributed to the Book trade worldwide by Springer, 2016.
9. T. Joachims, “A statistical learning model of text classification for support vector machines,” in *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 128–136, ACM Press, 2001.
10. F. A. Alshuwaier, W. A. Almutairi, and A. M. Areshey, “Smart search tools using named entity recognition,” in *2013 International Conference on Information Technology and Applications*, pp. 304–311, Nov 2013.
11. R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.
12. H. Sun, Z. Liu, and L. Kong, “A document clustering method based on hierarchical algorithm with model clustering,” in *22nd International Conference on Advanced Information Networking and Applications Workshops (aina workshops 2008)*, pp. 1229–1233, March 2008.
13. F. P. Shah and V. Patel, “A review on feature selection and feature extraction for text classification,” in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2264–2268, March 2016.
14. “Seleniumhq browser automation.” Seleniumhq.[Online]. Avilable: http: //www.seleniumhq.org/. [Accessed: Sept. 14, 2017].
15. L. Ou-Yang, “Newspaper: Article scraping & curation.” Newspaper Readthedocs.[Online].Avilable: https://newspaper.readthedocs. io.[Accessed: Sept. 14, 2017].
16. Y. Gu, S. Shen, J. Wang, and J. U. Kim, “Application of nosql database mongodb,” in *2015 IEEE International Conference on Consumer Electronics - Taiwan*, pp. 158–159, June 2015.
17. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, 2009.
18. C.-W. Hsu, C.-C. Chang, and C.-J. Lin, “A practical guide to support vector classification,” tech. rep., National Taiwan University, Taipei 106, Taiwan, 2003.
19. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
20. “Upazilas of bangladesh.” *Wikipedia*. [Online].Available: https://en.wikipedia.org/w/index.php?title=Upazilas of Bangladesh& oldid=807578395.[Accessed: Sept. 14, 2017].